HUGO RODY VIANNA SILVA

DESTINO EVOLUTIVO DE GENES DUPLICADOS EM PLANTAS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S586d
2016

Silva, Hugo Rody Vianna, 1985-
Destino evolutivo de genes duplicados em plantas / Hugo Rody Vianna Silva. – Viçosa, MG, 2016.
v, 119f. : il. (algumas color.) ; 29 cm.

Orientador: Luiz Orlando de Oliveira.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Evolução molecular. 2. Genética vegetal. 3. Poliploidia. I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-graduação em Bioquímica Aplicada. II. Título.

CDD 22. ed. 572.838

HUGO RODY VIANNA SILVA

DESTINO EVOLUTIVO DE GENES DUPLICADOS EM PLANTAS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*.

APROVADA: 26 de fevereiro de 2016.

Juliana Lopes Rangel Fietto

Luciano Gomes Fietto

Maximiller Dal Bianco Lamas Costa

José Miguel Ortega

Luiz Orlando de Oliveira
(Orientador)

ÍNDICE

LISTA DE ABREVIATURAS E SIGLAS

AMOVA	Analysis of Molecular Variance
BWT	Borrows-Wheeler Transform
CDS	Coding-DNA sequences
FPKM	Fragments Per Kilobase of exon per Million mapped reads
GBH	Gene-balance Hypothesis
GO	Gene Ontology
Ks	Synonymous substitutions per synonymous site
metE	Cobalamin-independent methionine synthase
MKT	McDonald and Kreitman Test
PCC	Pearson's Correlation Coeficient
PP	Posterior Probability
QC	Control Quality
RFO	Raffinose Family Oligosaccharides
SSD	Small-Scale Duplication
TF	Transcription Factors
WGD	Whole-Genome Duplication

RESUMO

SILVA, Hugo Rody Vianna, D.Sc., Universidade Federal de Viçosa, fevereiro de 2016. **Destino evolutivo de genes duplicados em plantas**. Orientador: Luiz Orlando de Oliveira.

Plantas são organismos paleopoliplóides que sobreviveram ao deletério processo de duplicação do genoma. Após a duplicação, cópias de um mesmo gene (parálogos) evoluem de forma divergente devido ao relaxamento da seleção purificadora, tornando os genes duplicados fonte de diversificação evolutiva. No entanto, nem todas as categorias de genes são retidas nos genomas, ou perdidas, de forma aleatória. Vários modelos evolutivos têm sido propostos para explicar o destino evolutivo e retenção tendenciosa de genes duplicados, para consequentemente entender melhor a poliploidia. Entretanto, a aplicabilidade destes modelos têm sido pouco investigadas. Usamos abordagens de genômica comparativa e de filogenia molecular para investigar o destino evolutivo de genes duplicados em plantas. Inicialmente, inferimos sobre a ancestralidade de duas pequenas famílias de genes duplicados — metionina sintase e oligossacarídeos de rafinose — em genomas de nove espécies de plantas superiores para, então, caracterizar as forças evolutivas que moldaram os destinos evolutivos dos parálogos, usando dados de resequenciamento de 31 genomas de soja. Posteriormente, usamos dados genômicos de 25 espécies de plantas taxonomicamente divergentes para associar mecanismos de duplicação, categoria funcional do gene e idade de duplicação na retenção tendenciosa de genes duplicados. Nas duas famílias gênicas, cada parálogo evoluiu de forma divergente devido ao relaxamento da seleção purificadora, o que permitiu que mutações aleatórias com valor adaptativo fossem eventualmente fixadas por seleção positiva. No entanto, a seleção purificadora parece ter sido mais restritiva em ao menos um dos parálogos de cada família gênica, preservando função ancestral. Tanto a idade quanto o mecanismo de duplicação contribuíram para variação nos padrões de retenção de genes duplicados nos 25 genomas alvos deste estudo. O destino evolutivo e a retenção de genes duplicados nos genomas parecem ser moldados por peculiaridades inerentes a cada organismo poliplóide.

ABSTRACT

SILVA, Hugo Rody Vianna, D.Sc., Universidade Federal de Viçosa, February, 2016. **The fate of duplicate genes in plants**. Adviser: Luiz Orlando de Oliveira.

Plants are paleopolyploid organisms that survived the deleterious process of genome duplication. After duplication, copies of the same gene (paralogs) evolve in different ways due to the relaxation of purifying selection, making the duplicated genes the greatest source of evolutionary diversification. However, not all categories of genes are retained in the genomes, or lost, randomly. Several evolutionary models have been proposed to explain the evolutionary fate and biased retention of duplicate genes, to thereby better understand polyploidy. However, the applicability of these models has been ill defined. We used approaches of comparative genomics and molecular phylogeny to investigate the fate of duplicated genes in plants. Initially, we inferred about the ancestry of two small families of duplicated genes — methionine synthase and raffinose oligosaccharides — in the genomes of nine species of plants to then characterize the evolutionary forces that have shaped the evolutionary fate of paralogs, using resequencing data from 31 soybean genomes. Later, we used genomic data from 25 taxonomically different plant species to associate mechanisms of duplication, functional gene category and age of duplication to the biased retention of duplicate genes. In each of the two gene families, paralog evolved in different ways due to the relaxation of purifying selection, which allowed random mutations with adaptive value be fixed by positive selection. However, purifying selection seems to be more restrictive in at least one of paralogs of each gene family, preserving the ancestral function. Both the age and the mechanism of duplication contributed to variation in duplicate genes retention patterns in the 25 target genomes in this study. The fate and retention of duplicate genes in the genomes is apparently shaped by peculiarities inherent to each polyploid organism.

INTRODUÇÃO GERAL

INTRODUÇÃO GERAL

Genes duplicados são reconhecidos como a maior fonte de inovação evolutiva (Lynch and Conery 2000; Freeling and Thomas 2006; Li et al. 2015). Apesar de as primeiras discussões sobre genes duplicados datarem da década de 1930, com os cientistas Hermann Joseph Muller e John Burdon Sanderson Haldane (Haldane 1932), somente em 1970 o assunto foi abordado em grande profundidade por Susumu Ohno, em seu livro intitulado "*Evolution by Gene Duplication*". Ohno concluiu que "*a duplicação de genes é único meio pelo qual um novo gene pode nascer*". Outra sugestão de Ohno foi a "Hipótese 2R", onde define os vertebrados como organismos paleopoliplóides, tendo sido originados devido a dois eventos (2 *rounds*) de duplicação de genoma inteiro (*whole-genome duplication* - WGD) em seu ancestral comum.

Em plantas, os sinais de poliploidia tem sido ainda melhor estudados. É possível que a maioria das espécies de plantas sejam na verdade seres paleopoliplóides. Algumas espécies como *Vitis vinifera*, por exemplo, tiveram seus genomas sequenciados mas as marcas de duplicações do genoma inteiro parecem ficar menos óbvias após milhares de anos de evolução (Jaillon et al. 2007). Em contrapartida, algumas espécies como *Glycine max* (Shoemaker et al. 2006), *Arabidopsis thaliana* (Bowers et al. 2003), e *Helianthus annuus* (Barker et al. 2008), possuem fortes sinais de terem sido acometidos a eventos recentes de poliploidia.

Subsequente à um evento de duplicação do genoma inteiro, o novo organismo poliplóide encontra diversos desafios. Além de superar sua inferioridade numérica no habitat em que vive, o poliplóide sofrerá severos rearranjos cromossômicos e deleções (Lynch and Conery 2000). Não tão somente, a diploidização tende a retornar o poliplóide para sua ploidia original. Assim, o genoma de cada organismo poliplóide não somente relata a história evolutiva daquele indivíduo mas também carrega informações valiosas para entender a poliploidia.

Nas últimas duas décadas, principalmente após os avanços das tecnologias de sequenciamento, vários autores tem usado genômica comparativa para elaborar hipóteses, além das propostas por Ohno, objetivando explicar o destino evolutivo dos genes duplicados. Ohno acreditava

que apenas uma cópia (parálogo) de um gene seria suficiente para manter a quantidade de produtos necessários para o metabolismo do organismo. Assim os parálogos evoluiriam de maneiras divergentes devido ao relaxamento da seleção purificadora (remoção de alelos deletérios). De acordo com Ohno, se por meio de mutações aleatórias, devido ao relaxamento da seleção purificadora, um parálogo adquirir qualquer vantagem estrutural que garanta aumento do *fitness*, estas mutações poderiam ser fixadas através de forte seleção positiva. Essa aquisição de vantagem foi chamada de neofuncionalização. Por outro lado, se mutações aleatórias comprometem a estrutura de um parálogo, o destino mais provável seria a perda de função ou pseudogenização. Em um terceiro possível modelo evolutivo proposto por Ohno, ambos os parálogos aceitariam mutações não completamente deletérias e culminariam compartilhando a função ancestral que anteriormente era cumprida por apenas uma cópia do gene. Esse modelo foi denominado subfuncionalização.

Enquanto que sob os modelos de Ohno a seleção purificadora estaria relaxada nos diferentes parálogos pois a duplicação de genes não alteraria o *fitness* do organismo poliplóide imeditamente após a duplicação, outros autores têm argumentado e proposto modelos evolutivos onde ambos o mecanismo de duplicação e a categoria funcional dos genes são determinantes para a retenção ou perda de parálogos. Um dos modelos evolutivos mais discutidos e embasados em dados é a “Hipótese do Equilíbrio Gênico” (*Gene Balance Hypothesis*) (Veitia 2002; Bowers et al. 2003; Papp et al. 2003). De acordo com essa hipótese algumas categorias de genes (genes conectados), como fatores de transcrição, seriam somente retidos nos genomas se originados por eventos de duplicação em larga escala como a duplicação do genoma inteiro. Isso porque este tipo de duplicação propicia ao poliplóide um aumento equiparado do conteúdo gênico, não culminando em desequilíbrio estequiométrico de produtos que poderia ter como consequência a diminuição do fitness ou mesmo letalidade. Por outro lado, genes em que seus produtos atuam de forma mais independente (genes não conectados) teriam tendencia de serem retidos quando duplicados por eventos de pequena escala (ex. duplicações em tandem). Em *A. thaiana*, por exemplo, fatores de transcrição são preferencialmente retidos nos genomas quando duplicados por eventos de

duplicação do genoma inteiro, enquanto que tendenciosamente perdidos após duplicações em pequena escala (Blanc and Wolfe 2004; Maere et al. 2005; Freeling 2009). Controversamente, algumas espécies de Asteraceae apresentam retenção tendenciosa de genes em duplicatas relacionados a complexidade estrutural, enquanto que fatores de transcrição foram significativamente sub-representados quando duplicados por duplicações do genoma inteiro (Barker et al. 2008). Mais controvérsias envolvem a Hipótese do Equilíbrio Gênico. Essa hipótese não explica, por exemplo, o papel dos mecanismos de regulação da expressão gênica. Fatores epigenéticos inerentes a cada organismo também não são abordados por esta hipótese. Por fim, apesar de ser uma das hipóteses mais embasadas por estudos em sistemas eucariotos, ela não foi ainda extensivamente testada.

Além da hipótese supracitada, vários outros modelos evolutivos de genes duplicados têm sido propostos, como: benefício pelo aumento de dosagem (Kondrashov and Koonin 2004), e fluxo metabólico (Hudson et al. 2011). Apesar disso, a real aplicabilidade de cada uma dessas hipóteses propostas são muito pouco conhecidas (Innan and Kondrashov 2010).

O sequenciamento de genomas não somente propiciou o aquecimento nas discussões sobre genes duplicados, como também permitirá que o tema seja estudado de maneira mais profunda. O resequenciamento de genomas também deverá representar um avanço fundamental. Medir a divergência entre parálogos em vários genomas de um mesmo organismo e entre organismos diferentes irá aumentar o número de perguntas que poderão ser respondidas e irá permitir também que os modelos evolutivos propostos sejam efetivamente testados.

ESTRUTURA E OBJETIVOS DA TESE

Esta tese foi dividida em três capítulos:

O primeiro capítulo representa um estudo sobre o destino evolutivo de genes duplicados em plantas. Neste estudo, nós usamos dados genômicos de nove espécies de plantas e dados de resequenciamento de 31 acessos de soja, incluindo genótipos silvestres e cultivados. Como modelo, nós usamos a pequena família de genes da metionina sintase independente de cobalamina e

determinamos a origem dos parálogos e quais as forças evolutivas moldaram a evolução desta família de genes em soja.

No capítulo 2 nós usamos dados genômicos de sete espécies de plantas e de resequenciamento de soja para estudar o destino evolutivo de genes envolvidos na produção de oligossacarídeos da família da rafinose (RFO). Nós inferimos sobre a ancestralidade dos RFOs, avaliamos as forças envolvidas na evolução destes genes em soja e estabelecemos qual modelo evolutivo de genes duplicados melhor se adapta aos dados.

O capítulo 3 apresenta um estudo onde testamos a retenção tendenciosa de genes duplicados. Usamos dados genômicos de 25 espécies de plantas e genômica comparativa para determinar quais categorias funcionais de genes são consistentemente superrepresentadas em duplicatas naqueles genomas quando originadas por duplicações de genoma inteiro ou tandem. Reproduzimos os resultados já publicados na literatura para algumas espécies e que levaram ao concebimento da Hipótese do Equilíbrio Gênico. Então aplicamos nossa metodologia em todas as espécies alvos do estudo pra avaliar se a variação observada na retenção de genes depende exclusivamente do mecanismo de duplicação ou da idade daquela duplicação.

REFERÊNCIAS

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* **25**: 2445–2455.

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. Unrevealing angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

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, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome*

Res **16**: 805–14.

Haldane JBS. 1932. *The Causes of Evolution*. Princeton University Press.

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.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**: 97–108.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Kondrashov FA, Koonin E V. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20**: 287–90.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Sci Adv* **1**: e1501084–e1501084.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science (80-)* **290**: 1151–1155.

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 U S A* **102**: 5454–5459.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.

Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Curr Opin Plant Biol* **9**: 104–9.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* **24**: 175–84.

Evolutionary history of a small gene family across the flowering plants: Insights from the cobalamin-independent methionine synthase

Evolutionary history of a small gene family across the flowering plants: Insights from the cobalamin-independent methionine synthase

Rody, Hugo Vianna Silva[a] and Oliveira, Luiz Orlando[a*]

[a]Department of Biochemistry and Molecular Biology, Federal University of Viçosa, 36570-900, MG, Brazil

***Corresponding author**
Luiz Orlando de Oliveira
Departamento de Bioquímica e Biologia Molecular
Av. P. H. Rolfs s/n
Universidade Federal de Viçosa
36570-000 Viçosa (MG), Brasil
Tel.: 55 31 3899 2964
Email: lorlando@ufv.br

Keywords: gene duplication, polyploidy, methionine synthase

ABSTRACT

Plants are well-succeed paleopolyploid organisms that increased in diversity by harboring sets of duplicate genes. Gene copies (paralogs) may have different fates depending upon the direction and intensities of several evolutionary forces that operate after the duplication. The mechanism by which a gene was duplicated and stoichiometric dosage constraints are also thought to play a crucial role in the immediate fixation and long term maintenance of paralogs in the genomes. Although much effort has been made in attempting to unravel the fate of duplicate genes, the proposed evolutionary models have been not tested in deep. In the study presented herein, we used genomic resources from nine phylogenetically distinct model plant species together with expression RNA-seq and Next-Generation Sequencing data from 31 genotypes of soybeans to investigate the evolution of duplicate genes, using the cobalamin-independent methionine synthase (metE) as a model gene. Our results showed metE as a small size gene family across flowering plants, encompassing a dual subcelullar distribution that arose very early on the evolutionary history of the angiosperms. One of the isoforms of metE harbors a transit peptide to chloroplast, while the other isoform remains in the cytoplasm. Soybeans presented paralogs from both isoforms that likely originated from polyploidy events. Were six soybean paralogs in the total, grouped into three methionine paralog pairs (MPPs). Out of these three pairs, the MPP1 holds the chloroplastic isoform and many of its amino acid sites have been subjected to positive selection. The two others are cytoplasm isoforms; one of which (MPP2) presented strong evidence for an ancient origin — they were located in the center of the genealogical network, were phylogetically related to single copy metE orthologs, and purifying selection was likely the main evolutionary force over these paralogs considering only few amino acid sites were predicted under positive selection. Our results support that the metE paralogs of soybeans are following the Ohno's neofunctionalization model of gene duplicate evolution. In short, the MPP2 maintained the amount of products required for the soybeans metabolism, while relaxed purifying selection allowed random mutations to be selected by positive selection in the other soybean metE paralogs.

INTRODUCTION

Polyploidy is certainly one of the most remarkable mechanism of duplication for promoting diversification of plant species (Simillion et al. 2002; Sémon and Wolfe 2007). Recent studies have shown that higher plants are in fact paleopolyploids; that is, they descend from an ancient polyploid ancestor (Bowers et al. 2003; Jaillon et al. 2007). The polyploids are intriguing; they survived the naturally deleterious process of chromosomes doubling, have overcome its numeric inferiority in the population, and exceed their parental species (Otto and Whitton 2000). After an event of whole-genome duplication (WGD), the resulting gene copies (paralogs) may have different fates depending upon the direction and intensities of several evolutionary forces that operate after the duplication (Freeling 2009; Edger and Pires 2009). According to Ohno's (1970) models, only one copy of a gene is enough to maintain a proper function of the organism, and therefore, the other paralogs would be under relaxed purifying selection, which in turn would favor neutral evolution and 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 become 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 Ohno, much effort has been made in an attempting to unravel the fate of duplicated gene pairs — especially after the recent advances in next-generation sequencing (NGS) technologies. Underlying biological attributes such as the mechanism of duplication (Papp et al. 2003; Birchler et al. 2001; 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 on the fate of gene duplicates and may corroborate for testing current gene models

within a genomic context. The cobalamin-independent methionine synthase (metE) gene, for example, represents an interesting model to investigate the fate of duplicated gene pairs. 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), Met plays a crucial role in many essential metabolic pathways as a donor of C1-units to SAM-dependent mythyltransferases (Hesse et al. 2004). In higher plants, the *de novo* synthesis of Met from the precursor *O*-phosphohomeserine requires only three enzymes: cystathionine γ-synthase, cystathionine β-lyase, and methionine synthase (MetE) (Ferrer et al. 2004). MetE (5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase; EC 2.1.1.14) catalyzes methionine 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 co-factor (Ferrer et al. 2004). MetE also takes part in the regeneration of the methyl group of the co-factor SAM after methylation reaction(Ferrer et al. 2004).

To date, the evolutionary history of metE gene family of higher plants remains an open question. In *Arabidopsis thaliana*, three variants of metE has been reported (Hesse et al. 2004; Ferrer et al. 2004) and doublets occurred in *Ammi majus* and *Petroselinum crispum* (Eichel et al. 1995). Other studies suggested that metE is a low copy gene in potato (Zeh et al. 2002) and soybeans (Hesse et al. 2004). Early studies suggest that *A. thaliana* possess MetE isoforms with a dual subcellular localization, that is, they are present in both the cytosol and the chloroplast (Ferrer et al. 2004). Crystal structures of MetE of higher plant species (*Arabidopsis thaliana*; Ferrer et al. 2004), thermophilic bacteria (*Thermotoga maritima*; (Pejchal and Ludwig 2004), and fungi (*Neurospora crassa*; (Wheatley et al. 2016) revealed striking structural and sequence similarities amongst these phylogenetically unrelated organisms: MetE is a monomeric protein with two domains, each containing a $(\alpha\beta)_8$ barrel.

Genome-wide sequence data of high quality from model plants species are available in public databanks, including Eudicots, Monocots and mosses. Moreover, the re-sequencing of several genomes of a single plant species, such as soybeans (*Glycine max*) (Lam et al. 2010) may contain crucial information to understanding the late evolution of duplicate gene pairs. Soybeans is a

paleopolyploid species; the soybean genome resulted from a hexaploidization event that took place at the origin of Eudicots (Jaillon et al. 2007) in addition to at least two WGD events of more recent origin (Shoemaker et al. 2006). About 4,500 years ago, soybeans was domesticated in China and since then this crop species has been under intense, artificial selection and plant breeding to combine superior agronomic traits. Due to the economic importance of soybeans, a collection of 31 accessions, including both wild and domesticated genotypes, had their entire genome re-sequenced (Lam et al. 2010). Finally, *Medicago truncatula*, which diverged from soybeans during one WGD event that took place about 44 million years ago (Pfeil et al. 2005), also had the genome sequenced (Young et al. 2011).

For the study reported herein, we investigated genomic resources from nine phylogenetically distinct, model plant species together with Next-Generation Sequencing data from the re-sequencing of 31 genotypes of soybeans to investigate the evolution of duplicate gene pairs of the metE gene family across the flowering plants. The following four questions were addressed: 1) How large is the metE gene family? 2) How ubiquitous is the dual subcellular distribution of MetE? 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?

MATERIAL AND METHODS

Assembling orthologs and paralogs of metE

Coding-DNA sequences (CDS) of soybeans, *Arabidopsis thaliana*, *A. lyrata*, *Lotus japonicus*, *Medicago truncatula*, *Vitis vinifera*, *Oryza sativa, Zea mays, Amborella trichopoda,* and *Physcomitrella patens* were obtained from PLAZA Dicots 3.0 (Proost et al. 2015). Gene annotations for soybeans 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 query in BLAST searches (Altschul et al. 1990) for orthologous genes against each of the nine CDS databases we had obtained from PLAZA. During searches, BLAST retained only primary non-redundant sequences that exhibited similarity

greater than 85% to the query sequence, with a minimum alignment length of 480bp — which is about 60% of the length of metE of *A. thaliana*. Finally, MUSCLE (Edgar 2004) aligned the sequences to create datasetA (N = 28; 2454 bp), which contained aligned CDS of metE orthologs across nine species of flowering plants. Additionally, DatasetB (N = 9; 2448 bp) included CDS from soybeans and *M. truncatula* only.

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

Bayesian phylogeny

We used Bayesian phylogeny to infer the phylogenetic relationships among orthologs and paralogs of metE. DatasetA was input into the MRMODELTEST v2 program (Nylander 2004) and the Akaike Information Criterion (Akaike 1973) 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 1 million generations each. The sequences of the orthologous genes of *P. patens* were used as outgroups. Trees were sampled once every 1000 generations. The selected settings ensured sufficient sampling of the posterior occurred; in Tracer 1.5 (Drummond et al. 2012), the Effective Sample Size values for the combined simultaneous runs were well above 500 for all statistics and the average standard deviation of split frequencies at the end of each run was below 0.01. 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 (PP) that were equal to bipartition frequencies. The consensus tree was visualized using FigTree v1.4 (http://tree.bio.ed.ac.uk/).

Divergence time estimation among orthologs and paralogs

The *yn00* tool from the PAML 4.1 package (Yang 2007) used datasetB to estimate the divergence times, in terms of synonymous substitutions per synonymous site (*Ks*), for all possible individual pairs of metE paralogs/orthologs between soybeans and *M. truncatula*. To infer whether the metE paralogs of soybeans originated through whole-genome duplication events, we confronted the congruence between the results of the divergence time estimation for the ortholog/paralogs pairs with the results of the Bayesian phylogeny. In soybeans, metE paralogs were considered as originated from whole-genome duplication events if they satisfied the following three criteria: (a) The rate between the number of paralogs of soybeans and their orthologs of *M. truncatula* was 2:1; (b) ortholog pairs of soybeans and *M. truncatula* shared one whole-genome duplication event (Shoemaker et al. 2006); and (c) the paralogs pairs of soybeans shared similar time of divergence.

Genomic data from wild and cultivated soybean genotypes

We downloaded NGS data from genomes of 31 re-sequenced soybean genotypes in paired-end reads with either 45-bp or 76-bp read length. The data were available at the National Center for Biotechnology Information (NCBI), under the accession number SRA020131. There were 17 genomes from wild genotypes and 14 genomes from domesticated genotypes. Expression RNA-seq data, protein-coding sequences, Coding-DNA sequences (CDS) and annotation files from the reference genome of soybeans (Wm82.a2.v1) were downloaded from Phytozome v10.3 (Goodstein et al. 2012).

After downloading NGS data, we used the makeblastdb tool available at NCBI to assemble local databanks for each of the 31 genomes. Subsequently, the BLASTn algorithm (Altschul et al. 1990) searched each genome for reads that matched each of the metE paralogs we had found on Wm82.a2.v1. Only reads that exhibited an e-value cutoff of e^{-05} were considered during subsequent analyses. The Quality Control (QC) was used to eliminate low-quality reads — that is, the reads that exhibited Q < 20 (McCormack et al. 2013). To have the sequence of each paralog assembled from sets of high-quality reads, we used

the *Borrows-Wheeler transform* (BWT) approach (Burrows and Wheeler 1994) in BOWTIE 2.0.5 (Langmead et al. 2009); this analysis used the metE paralogs of Wm82.a2.v1 as the reference sequences. The data had depth of coverage of about 10X.

Subsequently, we used the GATK package (McKenna et al. 2010) to inspect the aligned reads and to recalibrate the quality indices for each base. A total of three programs were used to estimate the presence of single nucleotide polymorphisms (SNP): Freebayes (Garrison and Marth 2012), an algorithm model based on Bayesian probability; VarScan (Koboldt et al. 2009); and SAMtools (Li et al. 2009). We declared a SNP when the three programs were in agreement, the locus had a minimum of 8X coverage depth, and the bases exhibited Q>20. Sequence regions harboring insertions/deletions (indels) or regions without coverage were trimmed off and discarded from subsequent analyses. A consensus sequence was assembled for each of the metE paralogs, in each of the 31 soybean genotypes. Supplementary Figure S1 shows a workflow depicting how the NGS data from genomes of 31 re-sequenced soybean genotypes gave rise to consensus sequences of metE paralogs.

Individual datasets were assembled for each of the metE paralogs of soybeans. The datasets contained metE sequences of each of 31 genotypes in addition to metE sequences of the soybean reference genome Wm82.a2.v1. Wm82.a2.v1 was taken as a domesticated genotype. MUSCLE aligned the sequences, which were finely edited in the Sequencher v4.8 program (Gene Codes Corp.). The metE paralogs of Wm82.a2.v1 were taken as the reference sequences; subsequently, we trimmed off and removed all non-coding sequences from those reference sequences. The “Automated Exploratory Recombination Analysis” tool available in the program RDP v3 (Martin et al. 2010) inspected our full dataset for intragenomic recombination events. Finally, we assembled DatasetC (N = 192; 2998 bp), which contained sequences of the metE paralogs of 17 wild genotypes and 15 domesticated genotypes.

Network analyses

Gene genealogies were inferred using the median-joining (MJ) network method (Bandelt et al. 1999) as implemented in NETWORK 4.6.1.4 (Fluxus Technology Ltd) with default parameters. Firstly, we inferred the genealogical relationships among the six paralogs of soybean using the full information of DatasetC. Next, we inferred the genealogical relationships among sequences from each of the six paralogs of soybean; thus, we split DatasetC into six subsets, each of which contained sequences from a given paralog.

Analyses of positive selection

To evaluate the hypothesis of positive selection considering the full extension of metE in soybeans, we applied the McDonald and Kreitman test (MKT) (McDonald and Kreitman 1997). The MKT test calculates the ratio of the number of non-synonymous polymorphic sites (*Pn*) by the number of synonyms polymorphic sites (*Ps*) within the species compared to the ratio of the number of non-synonymous nucleotide substitutions (*Dn*) by the number of synonymous nucleotide substitutions (*Ds*) between species; thus an outgroup is required to determine in which sites the differences are fixed (Bhatt et al. 2010). We prepared three version of datasetC. In each version, we added a set of sequences of metE orthologs from a distinct outgroup: *A. thaliana* (datasetC1; N = 195; 2448 bp), *M. truncatula* (datasetC2; N = 195; 2448 bp), and *O. sativa* (datasetC3; N = 194; 2448 bp). The MKT calculated the Neutrality Index (NI), which indicates how far polymorphism is from neutral evolution. If NI < 1, there is an excess of fixation of non-neutral substitutions and this indicates positive selection; NI > 1 suggests that negative selection is removing harmful mutations. Under neutral evolution, NI values are expected to be equal to 1. Additionally, the MKT calculated the rate of positive selection (α). This parameter can vary from -∞ to 1, where positive values of α corroborates with positive selection.

Adaptive changes have been suggested to occur only in a restricted set of protein sites (Bielawski and Yang 2003; Golding and Dean 1998), which complicates the efficient detection of positive selection at the level of whole

extent of a gene. Moreover, it has been demonstrated that the presence of slightly deleterious mutations may influence on the rate of positive selection (α) from polymorphism calculated by the MKT (Messer and Petrov 2013). Thus, we also used the *codeml* tool from the PAML 4.1 (Yang 2007) package to calculate the site-to-site ω variation, the ratio of nonsynonymous (d_N, amino acid changing) to synonymous (d_S, amino acid retaining) substitution rates ($\omega=d_N/d_S$) and the following codon-substituion models: M1a (neutral), M2a (selection), M7 (beta) and M8 (beta&ω). To test for positive selection in each of the metE paralogs of soybeans individually, datasetC was split into six sub-datasets: There was one sub-dataset for each paralog. To insure that ω variation represented amino acid sites that were fixed along independent lineages (Kryazhimskiy and Plotkin 2008), we incorporated the three metE ortholog sequences from *M. truncatula* to each of the sub-datasets (datasetC4 to datasetC10; N = 35). Possible ambiguities and alignment gaps were cleaned by PAML by choosing cleandata = 1. Additionally, unrooted neighbor-joining (NJ) phylogenetic trees were created using the Jukes-Cantor distance matrix model and used as input file. The likelihood ratio test (LRT) compared the models used; the positively selected codon sites were predicted using the Bayes Empirical Bayes (BEB) method. Codon sites were considered under positive selection when $\omega > 1$ and posterior probability calculated by the BEB was greater than 95%.

Visualization of the 3D structure of MetE was obtained in the PyMOL Viewer v1.5.0.1 (Schrödinger 2012), using as input the PDB file from *A. thaliana* available at the Protein Data Bank (PDB) under accession 1U1H. The protein structure was depicted as cartoon diagram and colored by chains.

Analyses of differential expression of paralogs

To investigate whether a given library would show the differential expression of a given metE paralogs of soybean, we recovered the Fragments Per Kilobase of exon per Million mapped reads (FPKM) values from the Phytozome's RNA-seq expression data files. Additionally, we constructed a Pearson's correlation coefficient (PCC) matrix to investigate co-expression of paralogs. This measure assumes that co-expressed genes follows a normal

distribution, where a coefficient equal to 1 means total positive correlation, 0 represents no correlation, and -1 indicates total negative correlation between the pair of paralogs being tested.

RESULTS

Genome-wide identification of orthologs and paralogs of metE

Searching the complete genome of nine species of flowering plants yielded a varying number of copies of metE per species. The reference genome of soybeans — Wm82.a2.v1 — yielded six copies, which displayed the following accession numbers: Glyma20G055900, Glyma05G090100, Glyma13G028400, Glyma16G038300, Glyma17G184900, and Glyma19G114500. We found three copies in *A. thaliana* (AT5g20980, AT3g03780, and AT5g17920), *A. lyrata* (AL3G03260, AL6G17860, and AL6G21270), *M. truncatula* (Medtr3g079640, Medtr6g027920, and Medtr7g086300), *Z. mays* (ZM01G29940, ZM01G48460, and ZM05G07210), and *P. patens* (PP00027G00280, PP00033G00570, and PP00399G00060); two copies in *Vitis vinifera* (GSVIVT01037511001 and GSVIVT01029971001), *A. trichopoda* (ATR_00004G00990 and ATR_00078G00850), and *O. sativa* (LOC_Os12g42876 and LOC_Os12g42884); and a single copy in *L. japonicus* (LJ1G054770). The “Automated Exploratory Recombination Analysis” tool inspected datasetC and suggested that recombination was mostly absent among the metE paralogs of soybeans. The only exception was a single recombination event that likely took place in Glyma16G038300 of the wild genotype W09.

Phylogeny of metE across the flowering plants

The Bayesian phylogenetic tree (Figure 1) showed well-supported nodes (PP = 100% on most of the nodes); it split the moss *P. patens* from the angiosperm clade. Within the angiosperm clade, there were two subclades; the first sub-clade comprised the Monocots (*O. sativa* and *Z. mays*) and the second sub-clade contained *A. trichopoda* together with the Eudicots. Nested within the sub-clade of the Eudicots, there were two third order-clades.

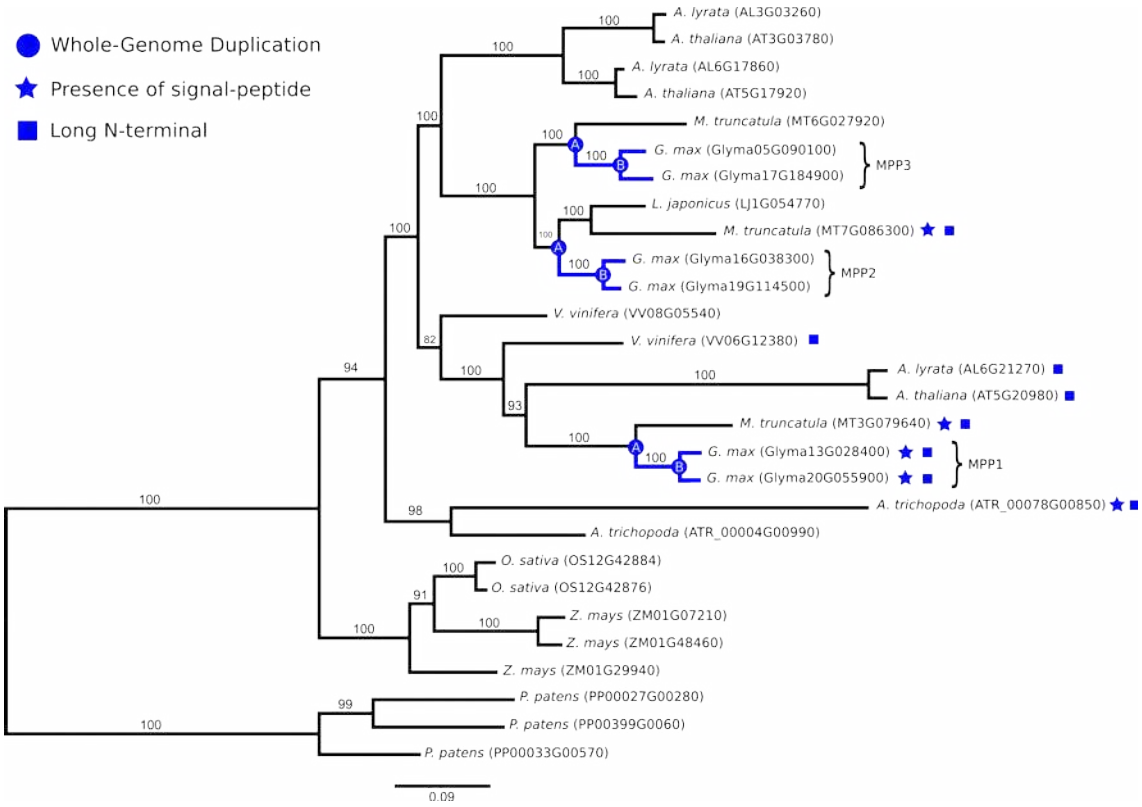


Figure 1. Bayesian phylogeny (consensus tree) showing the relationships among ortholog sequences of the metE gene family across the flowering plants. Orthologs of Physcomitrella patens were used as outgroups. Nodal support values are given as posterior probabilities (%) above the branches (when > 85%). Scale bar corresponds to the expected number of substitutions per site. Blue circles (labeled A and B) indicate the two whole-genome duplications of soybeans.

One third order-clades brought together metE sequences that lacked the chloroplast transit peptide in the N-terminal portion of the protein, while the other third order-clade contained sequences harboring the transit peptide. The six copies of metE of soybeans were split into three metE paralog pairs as follows: MPP1 (Glyma13G028400 and Glyma20G055900), MPP2 (Glyma19G114500 and Glyma16G038300), and MPP3 (Glyma05G090100 and Glyma17G184900). Each of the three metE paralog pairs of soybeans grouped together with one ortholog sequence of *M. truncatula* and formed a tip clade. Within each of those tip clades, the ratio of paralogs from soybeans to paralogs of *M. truncatula* was 2:1. The single copy ortholog of metE of *L. japonicus*

(LJ1G054770) grouped together with Medtr7g086300 and MPP2. Values of the *Ks* time divergence calculated for pairs of metE orthologs of soybeans and *M. truncatula* varied from 0.4106 to 0.5914 (Supplementary Table S2). We also calculated the *Ks* time divergence for the three closely related paralog pairs of soybeans; the pairwise *Ks* values varied from 0.1148 to 0.1589 (Supplementary Table S2).

Genealogical relationships among metE paralogs of soybeans

The genealogical relationships among the six metE paralogs of soybeans were clearly depicted on the resulting haplotype network (Figure 2A). The 192 sequences of DatasetC were grouped into six components, or groups of haplotypes. Each group of haplotypes consisted of sequences from a single paralog, without exception. Connections between the two nearest sequences of adjacent paralogs required a large number of mutational steps. The distances between members of distinct paralog pairs were larger than the distances between members of the same paralog pair. The largest number of mutational steps between members of distinct paralog pairs (470 steps) took place between a member of MPP1 (Glyma13G028400) and a member of MPP2 (Glyma16G038300); while the smallest number of mutational steps (63) occurred between members of MPP2 (Glyma19G114500 and Glyma16G038300). While MPP2 occupied the center of the network, MPP1 and MPP3 occupied each a tip position on opposite extremities of the network. Within each extremity, there was a member of the pair that diverged even further: Glyma20G055900 (from MPP1) and Glyma17G184900 (from MPP3). Additional networks investigated the genealogical relationships among haplotypes within a given paralog (Figure 2B to 2G). With few exceptions, each sequence gave rise to a haplotype. Within each intra-paralog network, the mutational distances that separated two nearest haplotypes were short.

The six resulting intra-paralog networks presented similar topologies; there was no apparent relationship between the position a given haplotype in the network and the origin of the haplotype — either from a domesticated genotype (coded blue) or a wild genotype (coded yellow).

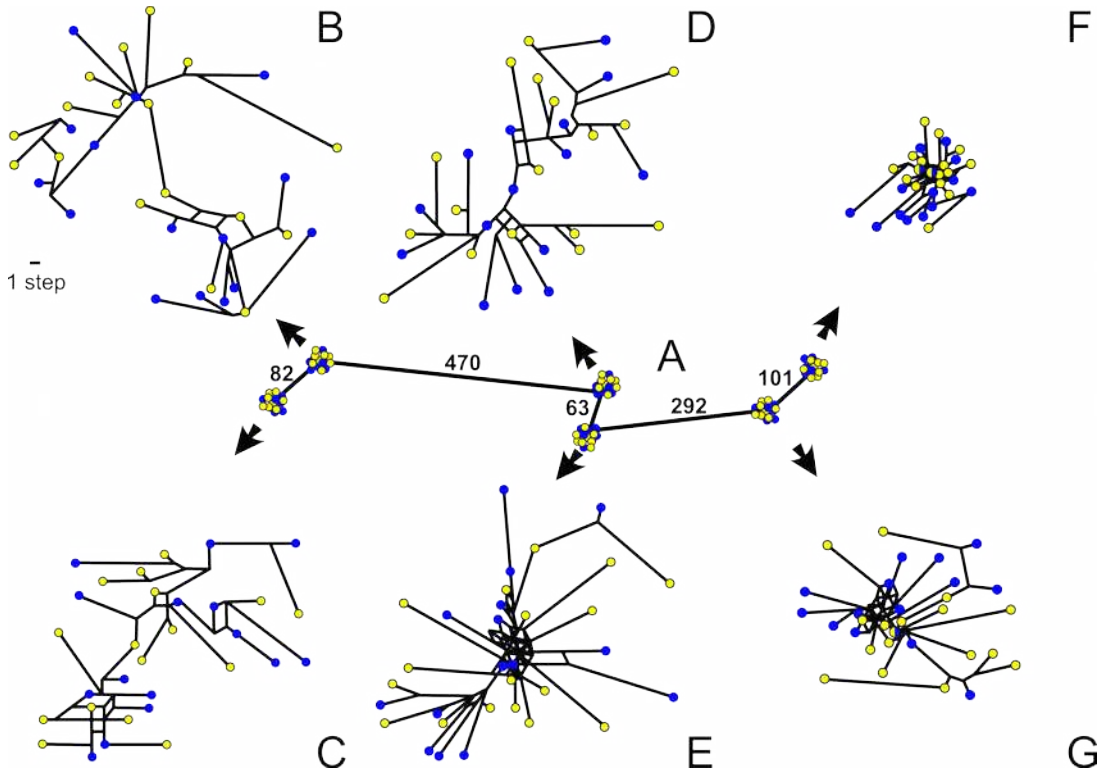


Figure 2. Median-joining networks for the six paralogs of metE of soybeans. Code for the networks are as follows: A, the full set of six paralogs; B and C (members of MPP1; Glyma13G028400 and Glyma20G055900); D and E (members of MPP2; Glyma19G114500 and Glyma16G038300); F and G (members of MPP3, Glyma05G090100 and Glyma17G184900). In A, the number of mutational steps is as indicated. From B to G, the length of the bar is proportional to the number of mutational steps. Color of cycles according to the origin of the genotypes: blue, domesticated; yellow, wild.

Positively selected sites on the metE paralogs of soybeans

In a preliminary attempt to evaluate positive selection, we employed MKT considering the full extension of the cDNA of each of the six metE paralogs of soybeans. Three independent MKTs were performed; each test used as outgroups orthologous sequences of the metE gene from a distinct species. These plant species displayed distinct phylogenetic relationships to soybeans: two sequences from *O. sativa* (Supplementary Table S3), three sequences from *A. thaliana* (Supplementary Table S4), and three sequences from *M. truncatula* (Supplementary Table S5). When orthologs of either *O. sativa* or *A. thaliana* were used as outgroup, results of the MKT showed NI < 1 for all metE paralogs

of soybeans. Those results were indicative of an excess of fixation of non-neutral substitutions on metE in soybeans, which suggested metE evolved under positive selection to a certain extent. When orthologs of *M. truncatula* were used as out-groups, NI values were > 1; thus, suggesting negative selection was removing harmful mutations.

Because it is unlikely that positive selection would impact all amino acid sites of a protein over a long period of time — as most protein amino acids evolve under strong constraints, we evaluated positive selection using site-to-site methods: M1a, M2a, M7, and M8 models implemented in the *codeml* tool available in the PAML 4.1 software package. Contrarily to the MKTs, which analyzed the full extension of the cDNA at once, those models analyzed the substitution rates ($\omega=d_N/d_S$) for each site of metE individually (Table 1). The results uncovered evidence for positively selected sites (LRT — $P < 0.01$; M1a and/or M7 rejected in favor of M2a and/or M8) in every paralog. The results obtained with the model M8 (Table 1) were shown because they were equivalent to those uncovered by the M2a model and because the M8 model detains more restrictive characteristics (Xu et al. 2013).

Table 1. Positively selected amino acid sites (ω > 1) obtained from coding-DNA sequence alignment of cobalamin-independent methionine synthase (metE) paralogs of 32 genotypes of soybeans. Prediction was performed by the codeml tool with model M8, implemented in PAML 4.1 software package and Bayesian posterior probabilities calculated by Bayes Empirical Bayes (BEB) >95% or >99%*.

Paralog pair	Paralogs	Positively selected amino acid position (amino acid)
MPP1	Glyma13G028400	40 (N), 252 (Y), 367 (V)*, 376 (T)*, 447 (R)*, 522 (V), 580 (G)*, 679 (V)*, 695 (K)*, 712 (S)*
	Glyma20G055900	40 (N), 268 (L), 367 (V), 376 (T)*, 447 (R), 495 (G)*, 521 (T), 522 (V), 580 (G)*, 659 (E)*, 679 (V)*, 695 (K)*
MPP2	Glyma19G114500	152 (V), 666 (R)
	Glyma16G038300	152 (V), 553 (R)
Mpp3	Glyma05G090100	131 (S)*, 512 (G), 618 (C)*, 659 (E)*, 680 (Y)*
	Glyma17G184900	131 (S)*, 512 (G)*

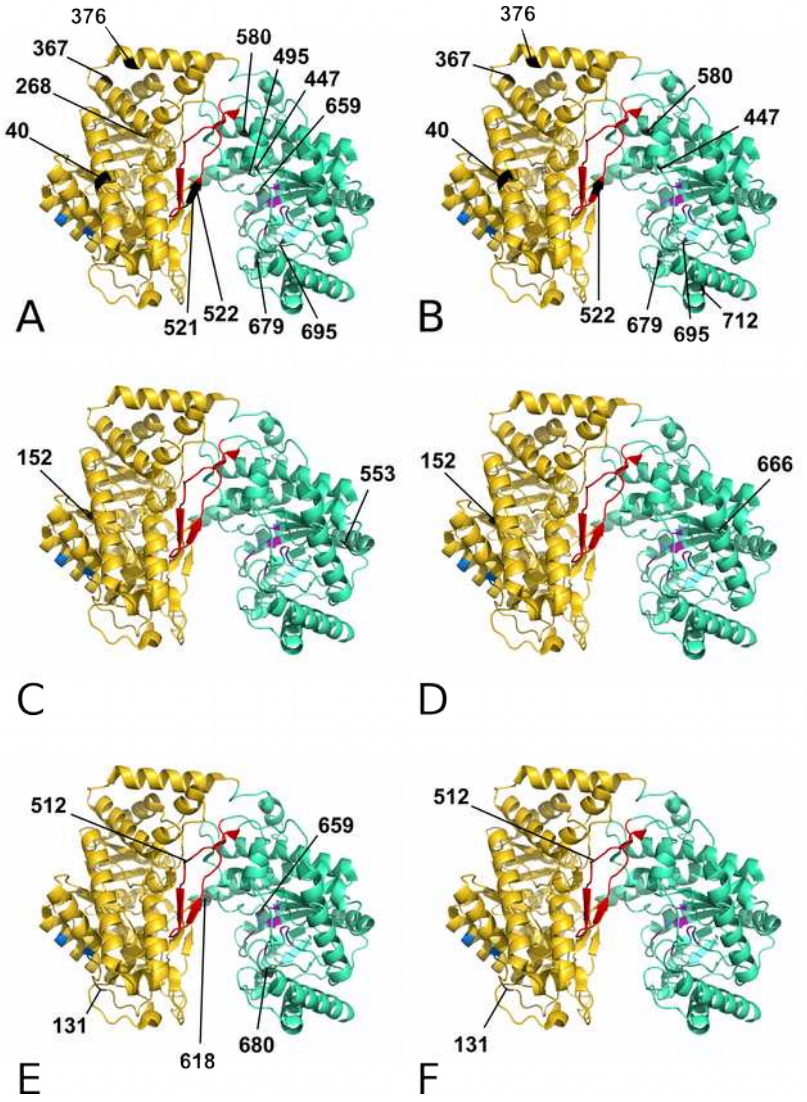


Figure 3. PyMOL diagrams depicting the 3D structure of the MetE of *A. thaliana* and the relative locations of positively selected sites of MetE of soybeans. Protein structure was taken from the Protein Data Bank (accession number 1U1H). Color codes: Yellow, the N-terminal domain; green cyan, the C-terminal domain; red, the loop involved in catalysis; black, the amino acid site selected by codeml/PAML 4.1 with Bayes Empirical Bayes posterior probabilities greater than 95% for each paralog of soybeans; purple, amino acid residues involved in the catalytic site (His^{647}, Cys^{649}, Cys^{733}, Asp^{605}, Ile^{437}, and Ser^{439}); and blue, the four zinc binding sites (His^{658}, Asp^{662}, His^{135}, and Asp^{194}). Letter codes: A and B (members of MPP1; Glyma13G028400 and Glyma20G055900); C and D (members of MPP2; Glyma19G114500 and Glyma16G038300); E and F (members of MPP3, Glyma05G090100 and Glyma17G184900).

A total of 21 sites were identified under positive selection with a posterior probability >95% as calculated by the BEB approach (Table 1). Most of the sites predicted to be under positive selection were restricted to a given paralog pair (Figure 3), with site 659 being the only exception (Figure 3A and 3C). Sites 521 and 522 were positively selected in MPP1 (Glyma20G055900 and Glyma13G028400), while site 512 was positively selected in MPP3 (Glyma05G090100 and Glyma17G184900). These three positively selected sites (512, 521, and 522) take part of a known loop region (Figure 3, depicted in red) of MetE; this loop is located between sites 507 and 529 (Ferrer et al. 2004) and interacts with the zinc atom at the C-terminal domain (sites 392 to 765). The number of positively selected sites varied among paralog pairs: Five sites on MPP1; three sites on MPP2, and 21 sites on MPP1.

Differential expression profile of metE paralogs of soybeans

Analyses of the RNA-seq expression data showed that metE paralogs of soybeans are differentially expressed (Figure 4). The Pearson's correlation coefficient matrix showed positive co-expression (PCC > 0.75) only between members of each methionine paralog pairs of soybeans (Supplementary Table S6). For example, the paralogs from MPP1 (Glyma13G028400 and Glyma20G055900) had PCC equal to 0.9374, but the two paralogs showed no co-expression with any other metE paralog. Because members within each paralog pair have high percentage of sequence identity (up to 85%) and owing to standard RNA-seq mapping issues such as multi-mapped reads, this expression in concert may be unrealistic.

Thus, we refrained from comparing the expression profile of isolated members within each the three paralog pairs and focused in showing the joint expression of members among pairs. In each of the eight libraries, MPP2 had a higher value of FPKM than either MPP1 or MPP3 (Figure 4, Supplementary Table S7). Despite the close phylogenetic relationship between MPP2 and MPP3 (see Figure 1), MPP3 always presented low values of FPKM (not exceeding 1.5, which is equivalent to zero) in libraries from either leaves or nodules. In most of the target libraries, MPP1 presented values of FPKM that were constantly low; the only exception was the seed library, in which FPKM

exhibited a small increment.

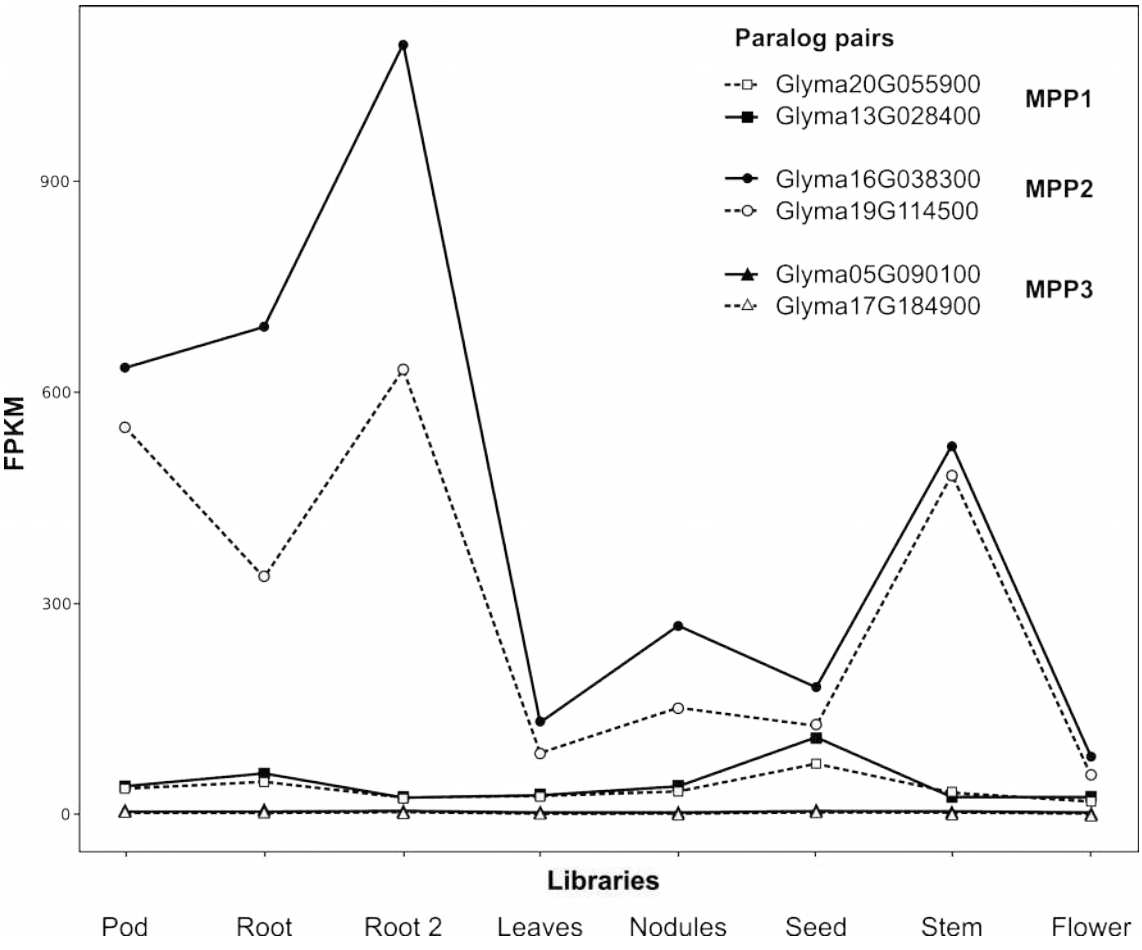


Figure 4. Expression profile of the six paralogs of metE of soybeans in eight RNA-seq libraries, disclosed in Fragments Per Kilobase of exon per Million mapped reads (FPKM) values.

DISCUSSION

The dual subcellular distribution of a small gene family

The small size of the metE gene family on taxonomically diverse plant species — one to six paralogs only — suggests that this feature may be of widespread occurrence across the land plants. Our genome survey revealed that the general pattern of the metE gene family in the Eudicots encompasses isoforms with a dual subcellular distribution: one of the isoforms was targeted to the chloroplast, while the other remained in the cytosol. The presence of both

the cytosolic and the chloroplastic isoforms of MetE on the most basal member of the angiosperms (*A. trichopoda*; Amborella Genome Project 2013) revealed that the dual subcellular distribution of MetE arose very early on the evolutionary history of the flowering plants and has been maintained over evolutionary timescales that encompass the diversification of all angiosperms. Previous studies have proposed distinct metabolic rules for each of those isoforms. The chloroplastic isoform of MetE would ensure *de novo* synthesis of Met from Hcy, therefore rendering these organelles autonomous for the synthesis of Met; while the cytosolic isoform would take part primarily in the regeneration the methyl group of SAM after transmethylation reaction (Ferrer et al. 2004).

Surprisingly, the dual subcellular distribution of MetE is absent from some extant species of flowering plants, such as the Eudicot *L. japonicus*, the Monocots *O. sativa, Z. mays*, and the moss *P. patens*. In those four plant species, MetE lacked transit peptides necessary to chloroplast targeting. It seems, therefore, that the autonomy of chloroplasts for *de novo* synthesis of Met is not as ubiquitous across the flowering plants as previously thought (Ferrer et al. 2004). Our sampling throughout the Monocots was restricted to two species only; thus, additional molecular analyses are necessary to confirm the extent of which chloroplast-targeted isoforms of MetE are absent from the Monocots.

WGD events gave rise to metE paralogs

We gathered strong, direct evidence for WGD as the likely origin of the metE paralogs of both soybeans and *M. truncatula*. In a 2:1 ratio, each of the three paralog pairs of soybeans (MPP1, MPP2, and MPP3) grouped together with one orthologous sequence of *M. truncatula*. Soybeans shares a common WGD with *M. truncatula*, which occurred around 44 million years ago (Pfeil et al. 2005). After the split, soybeans underwent a second, independent WGD event estimated to have occurred circa 14 mya (Shoemaker et al. 2006). Therefore, the 2:1 ratio confirmed that the metE paralogs of both soybeans and *M. truncatula* had their origin in WGD events. Based on the *Ks* time divergence between the soybeans-*Medicago* pair of orthologous, the older duplication

(labeled A in Figure 1) occurred around a *Ks* time equivalent of 0.5914. With a *Ks* time equivalent of 0.1589, the duplication that is exclusive of soybeans is much younger (labeled B in Figure 1) and resulted from the three parent paralogs giving rise to the six daughter paralogs that are currently present in the soybean genome. The likely origin of the gene duplicates we observed in the congeneric *A. thaliana* and *A lyrata* (in a 1:1 ratio) can be WGD events, which are known to have occurred in *Arabidopsis* (Blanc and Wolfe 2004).

The repeated episodes of large-scale deletions that accompany WGD events are known to reshape entire genomes, leading to the so called 'diploidization' of the neopolyploids (Conant et al. 2014). Through diploidization, the otherwise redundant copies are lost asymmetrically, while functional copies are maintained over time (Paterson et al. 2009). In *L. japonicus*, *O. sativa,* and *Z. mays*, discriminatory gene losses seems to have excluded from their genomes the metE paralogs that encoded the chloroplastic isoforms of MetE, while the metE paralogs of the cytosolic isoforms have been preserved. Whether *P. patens* never exhibited the dual subcellular distribution of MetE or lost its chloroplastic MetE remains an open question for the moment.

Ancestry of metE paralogs in soybeans

In soybeans, we uncovered three lines of evidence for an ancient origin of MPP2, one of the paralog pairs that encoded the cytosolic isoform of MetE. First, the topology of the full network (Figure 2A) showed the haplotypes MPP2 located in the center of the network, a position predicted by the Coalescent Theory (Templeton 1992) for lineages of parental status. Meanwhile, haplotypes of MPP1 (which encode the chloroplastic isoform of MetE) and haplotypes of MPP3 (which encode another cytosolic isoform of MetE) occupied each a tip position, which is consistent with a derived origin of these daughter copies (Templeton 1992). Second, the Bayesian phylogeny showed that MPP2, but not MPP3, displayed an orthologous relationship to LJ1G054770, the single copy ortholog of *L. japonicus*, as they grouped together with Medtr7g086300 (from *M. truncatula*) to form a well-supported sub-clade. Third, consistently with the predictions of the neofunctionalization model of duplicate genes (Ohno 1970; Walsh 1995) for parental copies, members of MPP2 were subjected to strong

purifying selection — the MetE isoform harbored a small number of positively selected sites (three sites only). The high levels of expression of both members of MPP2 across the eight target libraries (Figure 4) suggest that MPP2 maintains a defined biological function and that its MetE isoform is able to play a crucial role in Met metabolism across several tissues and organs. Our results, however, did not allow us to elaborate further about parental-daughter copy relationships between Glyma19G114500 and Glyma16G038300, the two members of MPP2.

Asymmetric functional divergence of metE paralogs of soybeans

The neofunctionalization model (Ohno 1970; Walsh 1995) postulates that purifying selection would maintain the original function of the parental copy; meanwhile releasing the daughter, redundant copies to acquire new functions in case the occurring mutations provided adaptive functional changes. Indeed, constraints of purifying selection were more relaxed over the daughter MPP1 and MPP3, which exhibited an accelerated rate of evolution and accumulated more sites predicted to be under positive selection then their parental MPP2.

The functional divergence of the ancestor ortholog of MPP1 included the acquisition of a chloroplast transit peptide, which most likely resulted from episodes of genome reshuffling that followed ancient duplication events during the diversification of the angiosperms. Most of the 21 positively selected sites of MPP1 do not correspond to amino acid residues in the enzyme's active site; thus, purifying selection played some role in maintaining this catalytic region of MPP1 free of changes (Figure 3). At the exclusion of MPP2, both MPP1 and MPP3 showed positively selected amino acid sites in the cationic loop region (residues 507 to 529), which is known to interact with the N-terminal domain and with a zinc atom at the C-terminal domain (Ferrer et al. 2004). The likely implications of the positively selected sites for attributing a new biological function the MetE isoform encoded by MPP1 and MPP3 remain elusive. The expression profile based on RNA-seq data clearly demonstrated differences on the product abundance of each the three MetE paralog pairs in different tissues of soybeans, with MPP3 exhibiting low FPKM values (not exceeding 1.5) in libraries from either leaves or nodules. Importantly, the MPP1 has increased

FPKM values in seed library, indicating that the translation of these paralogs may vary depending on intrinsic circumstances of the soybean metabolism.

Furthermore, our results do not fit on gene duplicate evolution models that argue for dosage constraints, either due increased dosage benefits (Kondrashov and Koonin 2004) or stoichiometric balance (Bowers et al. 2003; Blanc and Wolfe 2004a). To support these dosage hypothesis, we should have identified strong signals that all soybean MetE paralogs have resisted to alterations in its amino acid sites and we should have not observed such different expression profiles among paralogs. Thus, as only the members of one paralog pair (MPP2) were consistently expressed across all the soybean tissues, balance constraints seems to have not shaped the fate of MetE duplicate genes of soybeans. However, our results can be harmonized with the gene balance hypothesis, as proposed by Freeling (2008). As the soybeans MetE paralogs seems to have had their origin during WGD events, balance constraints may have been involved in the initial fase of fixation of MetE gene copies in the soybeans genome, which provided a long time frame for neofunctionalization.

ACKNOWLEDGMENTS

This study was supported by the Minas Gerais State Foundation of Research Aid - FAPEMIG (PPM 00561–15) to LOO. LOO received a fellowship from CNPq (PQ 304153/2012-5); HVSR received a fellowship from CAPES.

REFERENCES

Akaike H. 1973. *Information theory and an extension of maximum likelihood principle*. Second International Symposium on Information Theory, Budapest: Akademiai Kiado.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Amborella Genome Project. 2013. The Amborella genome and the evolution of flowering plants. *Science (80-)* **342**: 1241089.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.

Bhatt S, Katzourakis A, Pybus OG. 2010. Detecting natural selection in {RNA} virus populations using sequence summary statistics. *Infect Genet Evol* **10**: 421–430.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct genomics2* **3**: 201–212.

Birchler J a, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol* **234**: 275–288.

Blanc G, Wolfe KH. 2004a. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *Plant Cell* **16**: 1679–1691.

Blanc G, Wolfe KH. 2004b. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* **16**: 1667–1678.

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

Burrows M, Wheeler DJ. 1994. A Block-sorting Lossless Data Compression Algorithm. *Tech Rep* **124**.

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.

Corbi J, Debieu M, Rousselet A, Montalent P, Le Guilloux M, Manicacci D, Tenaillon MI. 2011. Contrasted patterns of selection since maize domestication on duplicated genes encoding a starch pathway enzyme. *Theor Appl Genet* **122**: 705–22.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**: 1969–73.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Edger PP, Pires JC. 2009. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosom Res* **17**: 699–717.

Eichel J, González JC, Hotze M, Matthews RG, Schröder J. 1995. Vitamin-B12-independent methionine synthase from a higher plant (Catharanthus roseus). Molecular characterization, regulation, heterologous expression, and enzyme properties. *Eur J Biochem* **230**: 1053–8.

Emanuelsson O, Brunak S, Heijne G von, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat Protoc* **2**: 953–971.

Ferrer J-L, Ravanel S, Robert M, Dumas R. 2004. Crystal Structures of Cobalamin-independent Methionine Synthase Complexed with Zinc, Homocysteine, and Methyltetrahydrofolate. *J Biol Chem* **279**: 44235–44238.

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. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn* **4**: 25–40.

Garrison E, Marth G. 2012. Freebayes. http://arxiv.org/abs/1207.3907.

Golding GB, Dean AM. 1998. The Structural Basis of Molecular Adaptation. *Mol Biol Evol* **15**: 355–369.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**:

Hesse H, Kreft O, Maimann S, Zeh M, Hoefgen R. 2004. Current understanding of the regulation of methionine biosynthesis in plants. *J Exp Bot* **55**: 1799–808.

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.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283–5.

Kondrashov FA, Koonin E V. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20**: 287–90.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet* **4**: e1000304.

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* **42**: 1053–9.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**: 2462–2463.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* **66**: 526–538.

McDonald JH, Kreitman M. 1997. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–654. http://dx.doi.org/10.1038/351652a0.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–303.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl Acad Sci U S A* **110**: 8615–20.

Nylander JAA. 2004. MrModeltest v2.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401–437.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**: 551–6.

Pejchal R, Ludwig ML. 2004. Cobalamin-Independent Methionine Synthase (MetE): A Face-to-Face Double Barrel That Evolved by Gene Duplication. *PLoS Biol* **3**: e31.

Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing Paleopolyploidy in Relation to Taxon Divergence: A Phylogenetic Analysis in Legumes Using 39 Gene Families. *Syst Biol* **54**: 441–454.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. 2015. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res* **43**: D974–81.

Schrödinger LCC. 2012. The PyMOL Molecular Graphics System.

www.pymol.org.

Sémon M, Wolfe KH. 2007. Consequences of genome duplication. *Curr Opin Genet Dev* **17**: 505–12.

Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Curr Opin Plant Biol* **9**: 104–9.

Simillion C, Vandepoele K, Montagu MCE Van, Zabeau M, Peer Y Van De. 2002. The hidden duplication past of Arabidopsis thaliana. *PNAS* **99**: 13627–13632.

Templeton AR. 1992. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**: 737.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* **24**: 175–84.

Walsh JB. 1995. How often do duplicated genes evolve new functions? *Genetics* **139**: 421–8.

Wheatley RW, Ng KKS, Kapoor M. 2016. Fungal cobalamin-independent methionine synthase: Insights from the model organism, Neurospora crassa. *Arch Biochem Biophys* **590**: 125–37.

Xu Z, Wu G, Li F, Bai J, Xing W, Zhang D, Zeng C. 2013. Positive selection signals of hepatitis B virus and their association with disease stages and viral genotypes. *Infect Genet Evol* **19**: 176–187.

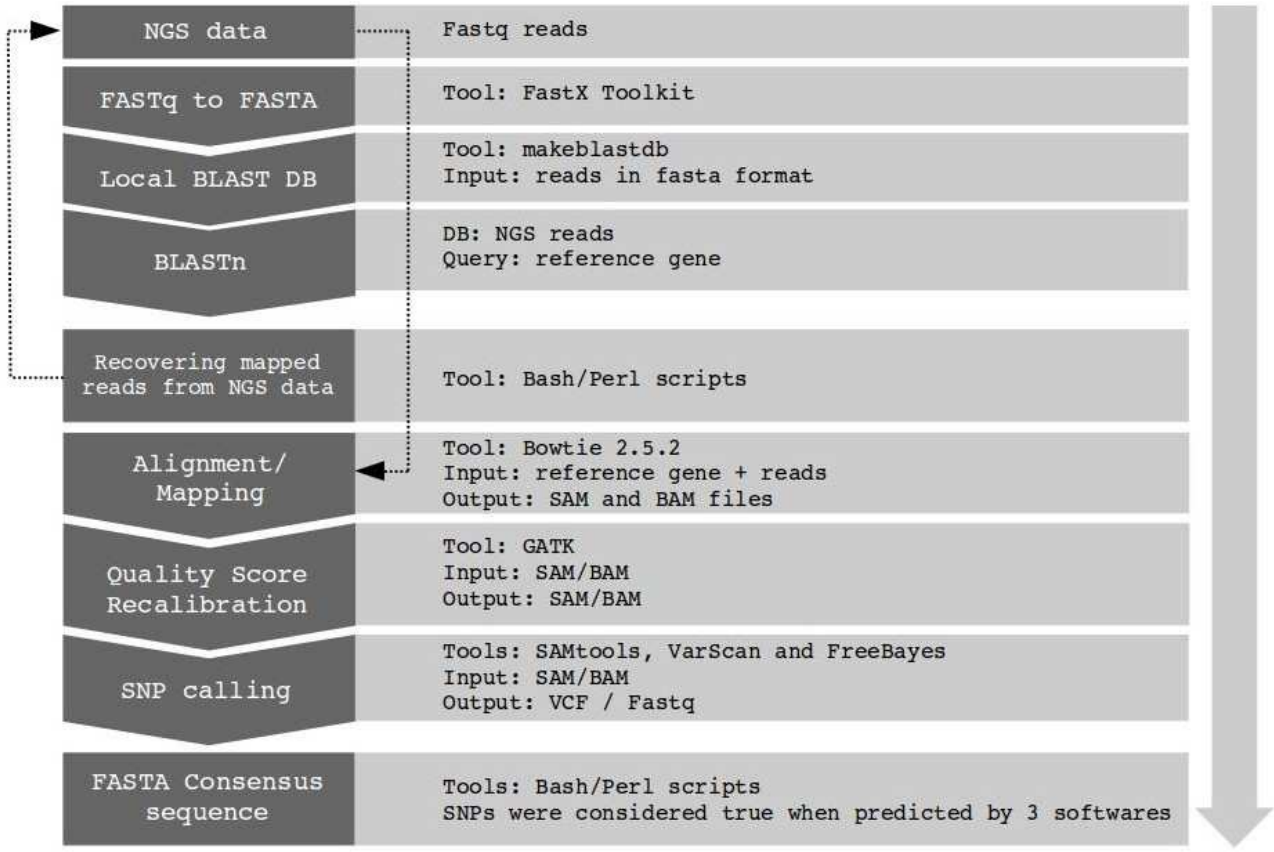
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, et al. 2011. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**: 520–4.

Zeh M, Leggewie G, Hoefgen R, Hesse H. 2002. Cloning and characterization of a cDNA encoding a cobalamin-independent methionine synthase from potato (Solanum tuberosum L.). *Plant Mol Biol* **48**: 255–65.

Evolutionary history of a small gene family across the flowering plants: Insights from the cobalamin-independent methionine synthase

Supplementary Material



Supplementary Figure S1. Workflow of the Next-generation sequencing data manipulation and SNP calling pipeline for the consensus sequence obtention.

Supplementary Table S1. Prediction of subcellular location performed by the TargetP v1.1 for 25 cobalamin-independent methionine synthase (MetE) ortholog sequences from nine flowering plant species.

Ortholog	Sequence length	cTP	mTP	SP	Other	Location	RC	TPlen
AL3G03260	2298	0.279	0.091	0.122	0.454	*	5	-
AL6G17860	2298	0.269	0.103	0.117	0.452	*	5	-
AL6G21270^{N}	2433	0.097	0.026	0.175	0.286	*	5	-
AT3G03780	2298	0.271	0.09	0.121	0.466	*	5	-
AT5G17920	2298	0.211	0.076	0.194	0.412	*	4	-
AT5G20980^{N}	2439	0.136	0.033	0.095	0.404	*	4	-
ATR_00004G00990	2298	0.287	0.061	0.177	0.443	*	5	-
ATR_00078G00850^{N}	2415	0.477	0.097	0.044	0.194	C	4	30
MT3G079640^{N}	2442	0.357	0.048	0.035	0.199	C	5	72
MT6G027920	2298	0.253	0.082	0.094	0.579	*	4	-
MT7G086300^{N}	2412	0.237	0.011	0.696	0.052	S	3	21
LJ1G054770	2292	0.108	0.18	0.087	0.659	*	3	-
Glyma20G055900^{N}	2400	0.412	0.029	0.073	0.21	C	4	22
Glyma05G090100	2298	0.14	0.131	0.173	0.555	*	4	-
Glyma13G028400^{N}	2412	0.386	0.024	0.142	0.158	C	4	22
Glyma16G038300	2292	0.199	0.055	0.31	0.384	*	5	-
Glyma17G184900	2298	0.149	0.127	0.167	0.567	*	4	-
Glyma19G114500	2292	0.14	0.044	0.495	0.326	S	5	23
OS12G42876	2301	0.123	0.116	0.165	0.624	*	3	-
OS12G42884	2301	0.106	0.091	0.286	0.479	*	5	-
VV06G12380^{N}	2415	0.321	0.04	0.008	0.479	*	5	-
VV08G05540	2010	0.242	0.045	0.062	0.465	*	4	-
ZM01G29940	2301	0.175	0.074	0.253	0.413	*	5	-
ZM01G48460	2298	0.154	0.107	0.142	0.581	*	3	-
ZM05G07210	2298	0.189	0.096	0.156	0.534	*	4	-

^{N} – MetE ortholog present longer N-terminal portion
CTP, mTP, SP, Other – Final scores on which the final prediction is based
C – the sequence contains cTP, a chloroplast transit peptide;
S – the sequence contains SP, a signal peptide;
* – any other location;
Tplen – Predicted presequence length.

Supplementary Table S2. *Ks* time divergence between paralog/ortholog pair sequences of cobalamin-independent methionine synthase (MetE).

Paralog/Ortholog pair	*Ks*
Gm20G055900 and Medtr3g079640	0.4106
Gm13G028400 and Medtr3g079640	0.4293
Gm05G090100 and Medtr6g027920	0.4231
Gm17G184900 and Medtr6g027920	0.4747
Gm19G114500 and Medtr7g086300	0.5871
Gm16G038300 and Medtr7g086300	0.5914
Gm20G055900 and Gm13G028400	0.1257
Gm05G090100 and Gm17G184900	0.1589
Gm16G038300 and Gm19G114500	0.1148

Supplementary Table S3. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of cobalamin-independent methionine synthase (MetE) paralogs of 18 wild and 14 domesticated soybeans, using two ortholog sequences of *O. sativa* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	α value	*P*-value
	Fixed differences between species (Ds)	Polymorphic sites (Ps)	Fixed differences between species (Dr)	Polymorphic sites (Pr)			
Glyma20G055900	455	65	715	46	0.450	0.550	0.000
Glyma13G028400	449	66	712	47	0.449	0.551	0.000
Glyma05G090100	406	73	723	41	0.315	0.685	0.000
Glyma17G184900	413	60	723	34	0324	0.676	0.000
Glyma16G038300	389	71	701	38	0.297	0.703	0.000
Glyma19G114500	393	78	700	40	0.288	0.712	0.000

Supplementary Table S4. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of cobalamin-independent methionine synthase (MetE) paralogs of 18 wild and 14 domesticated soybeans, using three ortholog sequences of *A. thaliana* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	α value	*P*-value
	Fixed differences between species (Ds)	Polymorphic sites (Ps)	Fixed differences between species (Dr)	Polymorphic sites (Pr)			
Glyma20G055900	275	508	636	204	0.174	0.826	0.000
Glyma13G028400	269	508	634	209	0.175	0.825	0.000
Glyma05G090100	255	520	646	202	0.153	0.847	0.000
Glyma17G184900	257	508	650	193	0.150	0.850	0.000
Glyma16G038300	244	512	638	194	0.145	0.855	0.000
Glyma19G114500	248	521	638	195	0.145	0.855	0.000

Supplementary Table S5. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of cobalamin-independent methionine synthase (MetE) paralogs of 18 wild and 14 domesticated soybeans, using three ortholog sequences of *M. truncatula* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	α value	*P*-value
	Fixed differences between species (Ds)	Polymorphic sites (Ps)	Fixed differences between species (Dr)	Polymorphic sites (Pr)			
Glyma20G055900	90	344	25	176	1.842	-0.842	0.010
Glyma13G028400	88	345	25	178	1.816	-0.816	0.014
Glyma05G090100	129	351	58	170	1.077	-0.077	0.716
Glyma17G184900	150	342	59	166	1.234	-0.234	0.251
Glyma16G038300	133	347	32	163	1.952	-0.952	0.002
Glyma19G114500	128	350	33	169	1.873	-0.873	0.003

Supplementary Table S6. Pearson's correlation coeficient (PCC) matrix of Fragments Per Kilobase of exon per Million mapped reads (FPKM) values of six cobalamin-independent methionine synthase (MetE) paralogs of soybeans. Significant co-expressed genes are given by PCC > 0.75*.

	Glyma05G090100	**Glyma17G184900**	**Glyma13G028400**	**Glyma20G055900**	**Glyma19G114500**	**Glyma16G038300**
Glyma05G090100	0	0.77609426*	-0.12911643	-0.39736451	-0.2809322	-0.36615055
Glyma17G184900	0.7760943*	0	0.08882896	-0.19643623	-0.1635826	-0.28133186
Glyma13G028400	-0.1291164	0.08882896	0	0.93743295*	-0.2745311	-0.24806638
Glyma20G055900	-0.3973645	-0.19643623	0.93743295*	0	-0.1053992	-0.08903299
Glyma19G114500	-0.2809322	-0.16358256	-0.27453105	-0.10539922	0	0.91665498*
Glyma16G038300	-0.3661506	-0.28133186	-0.24806638	-0.08903299	0.916655*	0

Supplementary Table S7. Expression profile of six cobalamin-independent methionine synthase (MetE) paralogs of soybeans in nine RNA-seq libraries, disclosed in Fragments Per Kilobase of exon per Million mapped reads (FPKM) values.

Paralog	Libraries							
	POD	Root	Root2	Leaves	Nodules	Seed	Stem	Flower
Glyma05G090100	0.0948448	0.0307723	0.0403859	0	0	0.00636102	0.0140393	0.411502
Glyma17G184900	0.0211194	0.0165281	0.0258517	0	0.0226241	0.00388295	0.0274272	0.229553
Glyma13G028400	37.566	55.115	22.116	25.9456	35.2716	112.031	24.7001	20.8917
Glyma20G055900	38.6577	45.474	22.4246	25.9244	31.3029	71.0006	29.7871	17.2625
Glyma19G114500	547.444	334.723	635.395	86.3252	149.022	127.197	481.406	53.2022
Glyma16G038300	637.016	692.894	1095.09	130.262	268.164	178.339	523.522	83.5427

Analysis of raffinose family oligosaccharide orthologs in flowering plants re-sheds light on Ohno's gene duplicate evolution model

Analysis of raffinose family oligosaccharide orthologs in flowering plants re-sheds light on Ohno's gene duplicate evolution model

Rody, Hugo Vianna Silva[a]; Guimarães, Valéria Monteze[a]; and Oliveira, Luiz Orlando[a*]

[a]Department of Biochemistry and Molecular Biology, Federal University of Viçosa, 36570-900, MG, Brazil

***Corresponding author**
Luiz Orlando de Oliveira
Departamento de Bioquímica e Biologia Molecular
Av. P. H. Rolfs s/n
Universidade Federal de Viçosa
36570-000 Viçosa (MG), Brasil
Tel.: 55 31 3899 2964
Email: lorlando@ufv.br

Keywords: duplicate genes, single copy genes, raffinose synthase

ABSTRACT

Gene duplications occurs in all extant eukaryotic organisms and are considered the key to understand the origin of genetic diversity. In plants, polyploidy is widely documented and have been increasingly along to advances in genomics. It is known that duplicate genes are tendentiously lost or retained in the genomes accordingly to features such as gene function, metabolic flux, mechanism of duplication, and age of duplication. Most of the proposed models were coined based in studies using measures of polymorphisms among all paralogs from a unique organism, which can sometimes be tricky and hide biases. Thus, it has become fundamental to deeply test the current gene duplicate evolution models. Here, we used a phylogenetic approach across flowering plants together with a genomic approach using resequencing data from 31 soybean accessions, including both wild and domesticated genotypes, to uncover the forces that have shaped the fate of genes from the raffinose family oligosaccharides (RFOs); a small family of genes of agronomic interest. We found that RFOs orthologs from seven flowering plants grouped into three distinct phylogenetic groups, each group harboring sequences from raffinose synthase, raffinose synthase related, and stachyose synthase enzymes. Monocots species differentiated from Eudicots as well as from the basal specie of *Amborella trichopoda* by lacking a amino acid insertion marker for the stachyose synthase enzyme. In soybeans, domestication seems to have not affected the RFOs, since the greatest genetic diversity was encountered among paralogs of soybeans. Strong signals of purifying selection was found for one gene copy within each of the three phylogenetic groups, while the remaining paralogs seemed to be under less evolutionary constraints which allowed random mutations occur in specific amino acid sites and be selected by positive selection. In the paralogs under purifying selection, only a few or no amino acid sites were predicted as positively selected. In general, the soybean RFO paralogs under purifying selection were higher expressed across eight RNA-seq libraries. All these results strongly corroborate with the neofunctionalization model of Ohno, on showing that for each gene function, one paralog were under strong evolutionary constraints while the remaining other paralogs were free to evolve.

INTRODUCTION

Gene duplication has long been thought as the main source of providing material for evolution (Haldane 1932; Ohno 1970; Jiao et al. 2011; Li et al. 2015). After either small- (tandem, segmental) or large-scale duplications (polyploidy), the gene copies (paralogs) basically undergo one of two possible fates: retention or loss. The mechanisms of duplication, functional gene category and the subsequent events after duplication are crucial for determining the fate of paralogs (Veitia 2002; Blanc and Wolfe 2004). Certain gene functional categories — transcription factors, for example — are preferentially retained when duplicated by whole-genome duplications (WGDs), but are disproportionately lost when originated from tandem events (Blanc and Wolfe 2004; Maere et al. 2005). Also as part of evolutionary process, chromosomal rearrangements and deletions after WGDs may dramatically change the genome content, as well as the fact that diploidization tends to return the polyploidy to its original condition. In general, dosage-sensitive genes — i.e., those that its product amount are stoichiometry required for proper organism metabolism — tend to be retained as duplicates because gene product imbalances would lead to disadvantageous fitness or lethality, while genes retained as single copy are dosage-insensitive (Edger and Pires 2009). On the other hand, Ohno's models of duplication gene evolution consider that duplication does not affect the organism fitness, at least in a short term after duplication, allowing that paralogs have different fates due to relaxed purifying selection (Ohno 1970). According to Ohno's models, a single gene copy would be sufficient to maintain the amount of products necessary for the organism metabolism, allowing the other gene copy to accumulate random mutations and diverge. For example, deleterious mutations could occur in a gene copy and make it lose its functionality (*pseudogenization*). Also, both copies could accumulate non-deleterious mutations and share the original function (*subfunctionalization*). Yet, if random mutations are advantageous and confer a new function for the gene (*neofunctionalization*), it would be fixed in the genome by positive selection.

Although many duplication gene evolution models have been proposed, their fundamental applicability and importance still remains apart. Innan and

Kondrashov (2010) elaborated a systematic classification guide of current understanding about gene duplicate evolution, and reinforced the needy to test the proposed models. For this purpose, they suggested comparative genomic approaches, combining expression data and sampling polymorphisms.

Measuring polymorphisms among paralogs in the whole gene content of an organism is often used to study patterns of gene duplication and retention (Maere et al. 2005; Hudson et al. 2011). This type of study is very important to understand the general trends of polyploidy. However, study duplicated genes in a small family of genes could provide greater focus to understanding the forces that shape polyploidy. At the same time, genes of agronomic interest could be further studied in crops. For example, due its extreme importance for world's economy, soybean had its genome full sequenced (Schmutz et al. 2010) and represent such a good crop organism to study polyploidy. In addition, soybean was resequenced several times (Lam et al. 2010), including both wild and domesticated genotypes, which allow us to determine if the measured polymorphisms tells the history of duplicated genes or only represents artificial selection. Although soybean is greatly consumed worldwide, the presence of oligosaccharides from raffinose family (RFOs) limits the consumption of this crop by constitute anti-nutritional components (Dierking and Bilyeu 2009). Therefore, the RFOs may represent a good gene family model to study gene duplicate evolution in soybean, since study this gene family may also represent advances in improving the nutritional characteristics of soybean.

Herein, we used data from seven species spread across the phylogeny of flowering plants and genomic data from the re-sequencing of 31 soybeans accessions, including both wild and domesticated genotypes, to study the fate of duplicate genes. A set of genes was selected from the raffinose family oligosaccharides (RFOs) to investigate the presence of selective forces based on polymorphism data. We also retrieved expression data from each of the paralogs to provide support for our study. In particular, we addressed the following objectives: 1) Infer on the ancestry of RFO orthologs; 2) Evaluate if domestication had influenced on the genetic diversity of soybean RFO paralogs; 3) Evaluate which evolutionary forces drive the fate of soybean RFO paralogs and establish which of the current duplicate gene evolution models best fit to the data.

METHODS

Determining RFO orthologs

Coding-DNA sequences and protein data from seven species — *Glycine max, Amborella trichopoda, Arabidopsis thaliana, Medicago truncatula, Vitis vinifera, Oryza sativa,* and *Zea mays* — were downloaded from PLAZA Dicots 3.0 (Proost et al. 2015). Gene annotation from soybean was downloaded from Phytozome v10.3 (Goodstein et al. 2012). A local BLAST database was created using the protein sequences and the *makeblastdb* tool made available by the National Center for Biotechnology Information (NCBI). The amino acid sequence of the raffinose synthase enzyme from *Arabidopsis thaliana* (ID AT5G40390) was used as query while searching for RFO homologous sequences in the local BLAST database, using the BLASTp algorithm with a cutoff of e^{-10}. All non-redundant sequences showing of at least 30% identity and 470 amino acid alignment length (representing at least 60% of the 784 amino acid length of AT5G40390) were considered as RFO orthologs. Finally, all retrieved sequences were aligned with MUSCLE (Edgar 2004) to create the DatasetA (N = 53; 3963 bp).

Bayesian phylogeny

We used Bayesian phylogeny to infer the phylogenetic relationships among orthologs of the RFO genes. Then, DatasetA was input into the MRMODELTEST v2 program (Nylander 2004) and the Akaike Information Criterion (Akaike 1973) 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 1 million generations each. One of the sequences of the orthologous genes of *A. trichopoda* was used as an outgroup. Trees were sampled once every 1000 generations. 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 that were equal to bipartition frequencies. The consensus tree was visualized using FigTree v1.4

(http://tree.bio.ed.ac.uk/).

NGS data manipulation: reconstructing RFO paralogs from resequenced soybean genomes

Genome resequencing data with either 45-bp or 76-bp paired-end read length from 31 soybean accessions were obtained from the NCBI, available under the accession number SRA020131. Out of the 31 soybean accessions, 14 were from domesticated soybean genotypes and 17 were wild genotypes. In order to reconstruct the sequences of soybean RFO paralogs, local databases were created independently for each of the 31 resequenced genomes using the makeblastdb tool. Subsequently, each soybean RFO paralog uncovered in the DatasetA was used as a query in the BLASTn searches to recover the reads from resequencing with a cutoff o fe^{-05}. The GATK package (McKenna et al. 2010) eliminated low quality reads (Quality Control - QC), determined by the minimum value *Q* > 20 (McCormack et al. 2013). The assembly of each sequence was performed by mapping the recovered reads for the corresponding soybean RFO paralog sequence, used as a reference. To this purpose, we used the *Borrows-Wheeler transform* (BWT) (Burrows and Wheeler 1994) approach implemented in the Bowtie 2.0.5 software (Langmead et al. 2009). From the aligned reads, the recalibration of quality indices was performed for each base using the GATK package (McKenna et al. 2010). Three programs were used to estimate the presence of single nucleotide polymorphisms (SNP): Freebayes (Garrison and Marth 2012), an algorithm model based on Bayesian probability; VarScan (Koboldt et al. 2009); and SAMtools (Li et al. 2009). We reported a SNP when the three programs were in agreement, the locus had a minimum of 8X coverage depth, and bases exhibited Q > 20. Insertions/deletions (indels) and sequence regions that had no coverage were trimmed and discarded from subsequent analyses. A consensus sequence was uncovered for each of the RFO paralogs in each of the 31 soybean genotypes. Sequences used as references were also incorporated to the analysis and considered as being from the wild soybean genotype. All sequences were finally aligned using MUSCLE (Edgar 2004) to create the DatasetB (N = 284; 2915 bp).

Genetic differentiation and neutrality tests

We firstly estimated genetic differentiation among the soybean RFO paralogs from wild and domesticated soybeans using the analyses of molecular variation (AMOVA) implemented in the software ARLEQUIN v3 (Excoffier et al. 2005). For this purpose, we split the DatasetB into two groups according to the origin of the soybean genotypes (wild or domesticated) and genetic differentiation was estimated using three hierarchical levels of variation: among paralogs, among groups within paralogs, and within paralogs. We also calculated the F-Statistics (Weir and Cockerham 1984) using ARLEQUIN to investigate the genetic partition among paralogs (F_{CT}), among groups within paralogs (F_{SC}), and within paralogs (F_{ST}). The significance of the genetic differentiation was tested with 1000 permutations, using a 95% confidence interval, where *P* shows the probability of having a more extreme variance component than those observed by chance alone. The F-values close to 0 indicate little or no difference between the groups and 1 indicates complete differentiation. We also used the ARLEQUIN v3 software to estimate measures of nucleotide diversity (number of segregating sites, S; and nucleotide diversity, π;) and to perform two tests of selective neutrality, Tajima's D (Tajima 1989) and Fu's FS (Fu 1997). Significant and negative values of D or F indicate an excess of low frequency polymorphism and supports the purifying selection hypothesis, while significant and positive values of D or F indicate that low frequency polymorphisms are lower than expected, supporting the balancing selection hypothesis. Non-significant values of D or F are consistent with the null hypothesis of neutrally evolving DNA.

Positive selection analysis

To evaluate the hypothesis of positive selection considering the entire extension of soybean RFO paralogs, we used the DatasetB and applied the McDonald and Kreitman test (MKT) (McDonald and Kreitman 1997) implemented in the software DNAsp v5.10.1 (Librado and Rozas 2009). The MKT test evaluates positive selection considering the ratio of the number of non-synonymous polymorphic sites (*Pn*) by the number of synonyms

polymorphic sites (*Ps*) within the species compared to the ratio of the number of non-synonymous nucleotide substitutions (*Dn*) by the number of synonymous nucleotide substitutions (*Ds*) between species, thus an outgroup sequence is required to determine in which sites the differences are fixed (Bhatt et al. 2010). Thus, only for performing the MKT is the DatasetB (N = 314; 3617 bp) modified by incorporating RFO orthologous sequences from four plant species from DatasetA: *A. trichopoda*, *O. sativa*, *A. thaliana*, and *M. truncatula*. The Neutrality Index (NI), calculated by the MKT indicates how far polymorphism is from neutral evolution. Thus, if NI < 1 there is excess fixation of non-neutral substitutions and this indicates positive selection. On the other hand, NI > 1 is expected when negative selection prevents harmful mutations. Under neutral evolution it is expected to find NI values equal to 1. Significance was indicated by the Fisher's exact test (two tailed) *P*-value < 0.05.

The program DATAMONKEY (Pond and Frost 2005) was used to evaluate a site-to-site positive selection in the DatasetB using three different methods: Single Likelihood Ancestor Counting (SLAC), Fixed Effects Likelihood (FEL) and Relaxed Effects Likelihood (REL), exploring three different approaches to evaluate a site-to-site positive selection. The SLAC is a method that counts the number of non-synonymous and synonymous substitutions along the phylogeny, the FEL estimates the ratio of non-synonymous to synonymous substitutions on a site-by-site basis and the REL assumes a distribution of rates across sites and infers the rate at which individual sites evolve based on this distribution. We used a 0.5 significance level for SLAC and FEL, and Bayes Factor up to 50 for REL to predict the selected amino acid sites. Use of the FEL test is indicated for our intermediate datasets (32 sequences for each soybean paralog), because of its performance advantage over the REL in terms of speed, and over the SLAC in terms of power as a function of the nominal α-level of the test (Pond and Frost 2005). Thus, it is important to note that the use of differential tests with different methods to estimate positive selection seems to be essential, especially when involving small data sets.

Analyses of differential expression of soybean RFO paralogs

To investigate the differential expression of soybean RFO paralogs in a given library, we recovered the Fragments Per Kilobase of exon per Million mapped reads (FPKM) values from the Phytozome's 10.3 (Goodstein et al. 2012) RNA-seq expression data files for *Glycine max* (Wm82.a2.v1). Additionally, a Pearson's correlation coefficient (PCC) matrix was constructed to investigate co-expression of paralogs. This measure assumes that co-expressed genes follow a normal distribution, where a coefficient equal to 1 means total positive correlation, 0 represents no correlation, and -1 indicates total negative correlation between the pair of paralogs being tested. Positive correlation was considered when the coefficient was greater than 0.75.

RESULTS AND DISCUSSION

The Bayesian phylogeny of RFO orthologs

Bayesian phylogeny was used to estimate relationships among the 53 RFO orthologs from the seven species of DatasetA. The Bayesian phylogenetic tree (Figure 1) showed well-supported nodes which split the RFO orthologs into three clades, herein referred to as: Group A (Figure 1, depicted in Blue), Group B (Figure 1, depicted in Green) and Group C (Figure 1, depicted in Orange).

We found the *A. thaliana* raffinose synthase ortholog (AT5G40390) nested within Group A. All the plant species used in this work presented RFO ortholog sequences in Group A. Soybeans had three RFO sequences — GM06G18890, GM05G02510, and GM05G08950 — in Group A, each grouped with one RFO ortholog from *M. truncatula* at a proportion of 1:1. These soybean RFO paralogs are annotated as raffinose synthase.

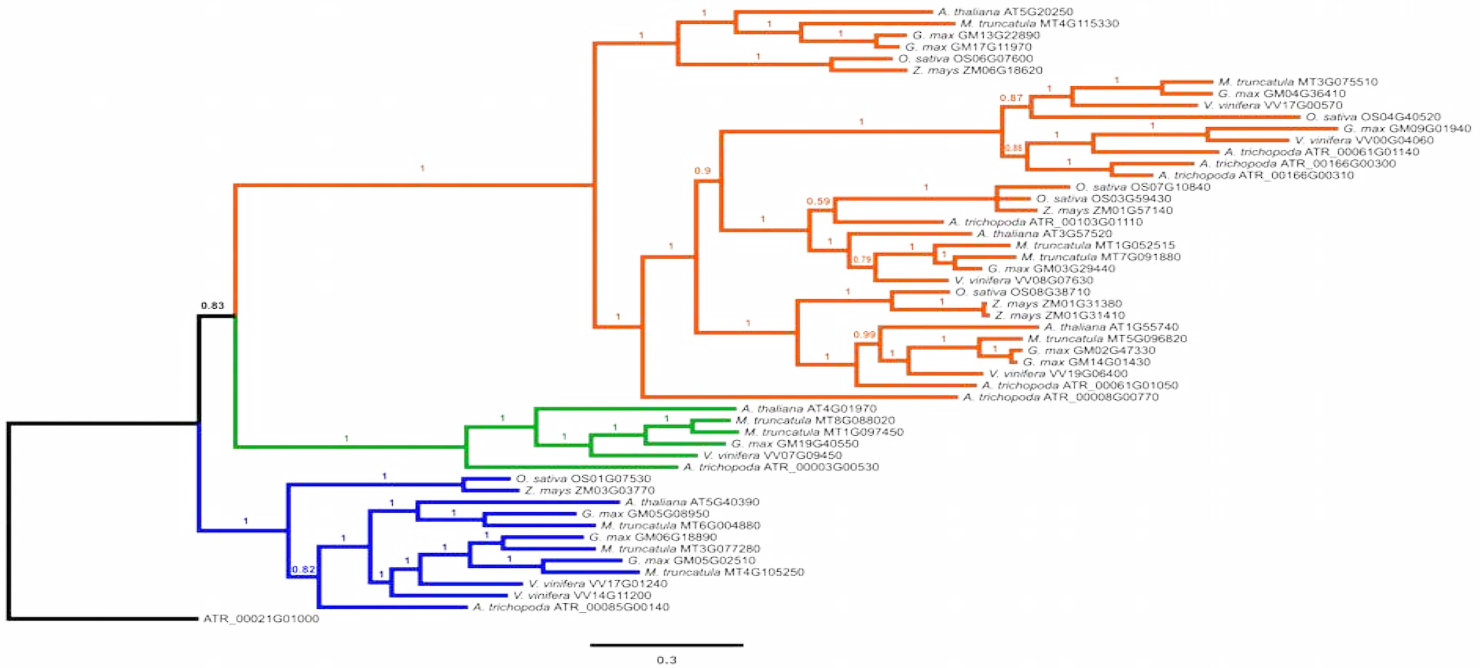


Figure 1. Bayesian phylogeny (consensus tree) showing the relationships among ortholog sequences of raffinose family oligosaccharides (RFO) from seven flowering plants (DatasetA). One sequence of A. trichopoda was used as the out-group. Nodal support values are given as posterior probabilities above the branches (when #85%). The scale bar corresponds to the expected number of substitutions per site. Colors indicate: Group A (blue) raffinose synthase orthologs; Group B (green) stachyose synthase orthologs; and Group C (orange) raffinose synthase related orthologs.

In Group B of the phylogenetic tree is harbored one RFO ortholog sequence from each Eudicot species, with the exception of *M. truncatula* which had two sequences in this group. The soybean RFO ortholog GM19G40550 is unique among soybean RFO paralogs, annotated as stachyose synthase (EC 2.4.1.67). The soybean GM19G40550 also differentiates from the other soybean RFO paralogs by having a insertion with approximately 80 amino acids (positions 304 to 383 of stachyose synthase from pea). This amino acid insertion appears to be a marker for the stachyose synthase protein in the plants we evaluated, since it is also present in other Eudicots such as *A. thaliana* (AT4G01970), *M. truncatula* (MT8G088020, MT1G097450), as well in the basal specie *A. trichopoda* (ATR_00003G00530). However, in the Monocots *O. sativa* and *Z. mays*, this insertion was not encountered. This fact is evidenced by the phylogenetic tree, where the two Monocot species evaluated had no sequences fitting in Group B. These data suggests that Monocots are destitute of this insertion marker for the stachyose synthase enzyme. Additionally, the soybean stachyose synthase protein GM19G40550 presented high similarity to its ortholog sequences from other Eudicots. For example, it is up to 75% similar to the pea (AJ311087) and 52% to *A. thaliana* (AT5G40390). This high similarity among RFO orthologs from pea to other plants has been previously observed (Peterbauer et al. 2002).

Soybeans and *M. truncatula* have a very similar history of poliploidy. These species share a recent whole-genome duplication event, where the soybean has undergone one additional and exclusive WGD event (Shoemaker et al. 2006) after its speciation. Considering the aforementioned, we would expect to observe a gene proportion of 2:1 (soybean:medicago) when comparing these two species in the case that there were no deletions in these species. Therefore, because this proportion seems to be inverted in the case of the stachyose synthase gene — *M. truncatula* has 2 while soybean has 1 —, and the proportion is 1:1 in the case of raffinose synthase, we hypothesize that successive fractionation occurred in these species, culminating in the loss of copies of the respective RFO paralogs during the course of millions of years of evolution.

In the Group C, the proportion of duplicate genes expected in the comparison between *G. max* and *M. truncatula* becomes closer to what would

be expected. For example, the Bayesian phylogenetic tree showed a close relationship among two soybean RFO sequences (GM14G01430 and GM02G47330) with one *M. truncatula* ortholog (MT5G09820). These results are an indication that not all but some soybean RFO paralogs have resisted fractionation and chromosome rearrangements that tend to return the polyploid to its fundamental diploid condition.

Domestication has not affected the soybean RFO paralogs

Out of the 11 soybean RFO paralog sequences from DatasetA, two sequences — GM13G22890 and GM17G11970 — were excluded from DatasetB and from the next analysis because of the low density of reads in the 31 resequenced soybean genomes. Thus, nine sequences were declared as soybean RFO paralogs and further investigated in this study.

To evaluate if domestication had influenced on the genetic diversity of soybean RFO paralogs, we used AMOVA (Table 1) in DatasetB. We ran one AMOVA with three hierarchical levels: Among paralogs, Among groups within paralogs — groups according to the wild or domesticated soybean genotype —, and within paralogs. The results showed that the greatest differentiation occurs significantly among paralogs (F_{CT} = 0.989, P < 0.01), and not among groups of wild or domesticated soybean genotypes (F_{SC} = 0.039, P = 0.04). These data are an indication that soybean RFO paralogs were not affected by artificial selection during the domestication process. Because of this result, each the soybean RFO paralog sequences from either wild or domesticated genotypes were evaluated together in further analysis.

Table 1. Analysis of Molecular Variance (AMOVA) of nine raffinose family oligosaccharide (RFO) paralogs of soybean, split into groups of either wild (n = 18) or domesticated (n = 14) genotypes, subdivided in three hierarchical levels.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	*P*-value
Among paralogs	8	155473.966	617.84704 Va	98.96	0.00
Among groups within paralogs	9	91.474	0.25461 Vb	0.04	0.04
Within paralogs	282	1646.811	6.21438 Vc	1.00	0.00

Conservation of ancestral gene function and divergent evolution due to relaxed purifying selection of soybean RFO paralogs

We found negative significant values of D, indicative of purifying selection, for three soybean RFO paralogs: GM05G02510, GM19G40550, and GM02G47330 (Table 2). Out of these three soybean RFO paralogs, only the paralog GM05G02510 is associated with the phylogenetic Group A (Figure 1, depicted in blue), the group with which the *A. thaliana* raffinose synthase (AT5G40390) is also related. This data may be an indication that this soybean paralog might play an important role in the production of raffinose oligosaccharides in soybean. However, Dierking and Bilyeu (2008) suggested that the soybean raffinose synthase paralog RS2 (GenBank id # EU651888) is the main agent responsible for the synthesis of raffinose in soybean; this paralog has 100% identity to the paralog GM06G18890. The second soybean RFO paralog found under purifying selection is the GM19G40550 stachyose synthase, which has its single copy status revealed by the Bayesian phylogeny (Figure 1, Group B, depicted in green).

Table 2. Measures of nucleotide diversities and neutrality test statistics from coding-DNA sequence alignment of raffinose family oligosaccharide (RFO) paralogs of soybeans.

Paralog	Sample Size	Number of segregating sites (S)	Nucleotide diversity (π)	Tajima's D	*P*-value	FS	*P*-value
GM02G47330	32	83	5.98992	-2.66885	0.00	-17.90649	0.00
GM03G29440	32	69	10.42339	-1.46427	0.06	-24.53943	0.00
GM04G36410	32	31	7.74597	0.02252	0.57	-24.04774	0.00
GM05G08950	32	23	4.08468	-0.99422	0.15	-25.47386	0.00
GM05G02510	32	14	1.08468	-2.26591	0.00	-11.4843	0.00
GM06G18890	32	19	3.30847	-1.02245	0.16	-25.91696	0.00
GM09G01940	32	22	3.27016	-1.39518	0.07	-19.87375	0.00
GM14G01430	32	40	8.14516	-0.65553	0.29	-23.4488	0.00
GM19G40550	28	8	0.80423	-1.87980	0.01	-6.76469	0.00

The third soybean RFO paralog (GM02G47330) under evolutionary constraints is associated with the phylogenetic Group C (Figure 1, depicted in orange), the group with the larger number of RFO orthologs. In summary, each of the three specific groups revealed by the Bayesian phylogeny (Figure 1) had only one soybean RFO paralog predicted as under purifying selection by the Tajima's D test. All the remaining soybean RFO paralogs showed no significant values of D, which is an indication of evolution under neutrality. These results are in accordance with the Ohno's neofunctionalization gene duplication evolution model (Ohno 1970), sustaining that after duplication one gene copy is sufficient to maintain the proper amount of the gene product required by the organism, while the other gene copies are under relaxed constraints and can therefore accumulate mutations or be lost. In *A. thaliana*, for example, it has already been demonstrated (Egert et al. 2013) that only one of the RFO paralogs (AT5G40390) is responsible for accumulation of raffinose synthase in the leaves under different stress conditions, such as cold, dry, salt, and thermal shock.

The MKT was used to evaluate positive selection considering the entire extension of the soybean RFO paralogs, using outgroup sequences from three plant species — *A. trichopoda*, *O. sativa, A. thaliana, and M. truncatula* — with a distinct evolutionary relationship to soybean (Supplementary Table S1, S2, S3, and S4).

No significant positive selection was uncovered by the MKT for any of soybean RFO paralogs when using orthologous sequences from the four species as an outgroup. However, when using orthologous sequences from the Monocot *O. sativa*, MKT revealed significant values ($P < 0.05$) of NI > 1, suggesting negative selection in four soybean RFO paralogs: GM03G29440, GM04G36410, GM09G01940, and GM14G01430. These paralogs have close phylogenetic relationships, and were all harbored in Group C of phylogenetic tree (Figure 1, depicted in orange). These results suggest that mutations in the soybean RFO paralogs from Group C are unlikely to occur and be fixed. Additionally, because negative selection was only revealed when using sequences from *O. sativa* as the outgroup, we argue that the most important changes in RFO protein structure may have occurred based on Monocots and Eudicots, being a more general pattern across flowering plants and not

restricted only to the soybean RFO paralogs.

Table 3. Sites under positive selection in paralogs from soybeans, predicted by the three methods available on DataMonkey. Single Likelihood Ancestor Counting (SLAC), and FEL methods, considered 0.5 of significance level. REL method, considered Bayes Factor > 50 as significance level.

Paralog	Method	Positively selected amino acid sites
GM02G47330		37, 87, 107, 343
GM03G29440		291, 307, 352, 358, 408, 456, 536, 611, 658, 676, 717
GM04G36410		119, 134, 141, 164, 192, 208, 266, 631
GM05G02510		none
GM05G08950	**SLAC**	16, 729
GM06G18890		693, 738
GM09G01940		451, 517
GM14G01430		31, 35, 87, 298, 339, 429, 572, 594, 603, 653
GM19G40550		none
GM02G47330		35, 37, 40, 42, 87, 90, 107, 308, 339, 343, 344, 352, 355, 357, 438, 440, 572, 585, 598, 656, 666, 679, 683
GM03G29440		291, 307, 338, 349, 351, 352, 358, 376, 384, 405, 408, 418, 422, 459, 477, 513, 536, 554, 611, 617, 658, 664, 717
GM04G36410		119, 134, 141, 164, 208, 250, 253, 266, 353, 388, 411, 429, 463, 575, 631, 632
GM05G02510	**FEL**	41, 66, 85, 172, 355, 425
GM05G08950		10, 16, 49, 655, 721, 722, 729
GM06G18890		667, 694, 708, 738
GM09G01940		87, 257, 417, 418, 451, 517, 724
GM14G01430		9, 13, 14, 31, 35, 87, 115, 224, 298, 339, 375, 429, 456, 594, 603, 653, 683, 687
GM19G40550		545, 634
GM02G47330		none
GM03G29440		307, 338, 349, 352, 422, 513, 536, 611, 717
GM04G36410		119, 134, 141, 164, 192, 208, 250, 253, 353, 388, 411, 429, 575, 631, 632
GM05G02510		none
GM05G08950	**REL**	16, 655, 729
GM06G18890		667, 694, 738
GM09G01940		43, 56, 87, 191, 257, 417, 418, 451, 478, 517, 724
GM14G01430		22, 31, 35, 298, 339, 560, 603
GM19G40550		none

Because positive selection may occur only in few sites of a protein due its structural concerns, we applied three more tests (Table 3) involving different statistical approaches to evaluate a site-to-site positive selection in the soybean RFO paralogs. Positively selected amino acid sites were predicted by all three SLAC, FEL, and REL methods for the RFO soybean paralogs. The most sensitive test was the FEL at a significance level of 0.5. For example, no positively selected sites were predicted by the SLAC and REL methods for the two soybean RFO paralogs GM05G02510 and GM19G40550. The FEL method, however, was able to predict six and two amino acid sites, respectively, under positive selections in these two soybean paralogs. In general, soybean paralogs that were early predicted as under neutral evolution (Table 2), and belonging to phylogenetic Group C, had a greater number of positively selected amino acid sites than those RFO paralogs predicted as under purifying selection. We speculate that the relaxation of negative selection facilitated the occurrence of mutations and signaled by positive selection, where even the majority of amino acid sites for these soybean RFO paralogs are unlikely to accept mutations. These results may partially explain why some of these RFO paralogs are retained in the soybean genome, since positive selection is commonly linked to fixation of gene duplicates (Innan and Kondrashov 2010).

No co-expression among the soybean RFO paralogs or signal of positive correlation was encountered, with coefficient PCC > 0.75 (Supplementary Table S5). We observed that the paralogs are differentially expressed across the eight libraries (Table 4), which suggests that not a unique RFO soybean paralog, but all may have a specific contribution in the production and maintenance of proper levels of the raffinose family oligosaccharides in the different soybean tissues.

Analyzing the expression profile of only the paralogs annotated as raffinose synthase from the phylogenetic Group A — GM05G08950, GM05G02510, and GM06G18890 —, we found the paralog GM05G08950 expressed along all eight RNA-seq libraries, although more expressed in the leaves. The paralogs GM05G02510 and GM06G18890, previously predicted as under purifying selection in this study (Table 2), were more expressed in the stem and seed libraries, respectively. Importantly, the GM06G18890 presented the highest value of FPKM among all the soybean RFO paralogs in the seed library. The accumulation of raffinose in the seeds is in accordance with the

hypothesis that these oligosaccharides serve as osmoprotectants and antioxidant compounds in developing seeds (Castillo et al. 1990; Li et al. 2011; Collakova et al. 2013).

Similarly, Dierking and Bilyeu (2008) also found the GM06G18890 paralog associated with the seed phenotype of soybeans when using quantitative RT-PCR. However, these last authors did not determined differences in the expression level of raffinose synthase candidate genes. The stachyose synthase paralog GM19G40550 did not present high values of FPKM when compared to the other soybean paralogs, but this paralog appears to be expressed in all the evaluated RNA-seq libraries, with elevated FPKM values in the stem and leave tissues.

Special attention should be given to the paralogs GM03G29440 and GM14G01430 from the phylogenetic Group C, annotated as related raffinose synthases. They were more expressed than all the other paralogs across the POD, root, nodules, and flower RNA-seq libraries. Interestingly, these last two paralogs were predicted by the MKT as under negative selection.

Table 4. Expression profile of nine raffinose family oligosaccharide (RFO) paralogs of soybeans in eight RNA-seq libraries, disclosed in Fragments Per Kilobase of exon per Million mapped reads (FPKM) values.

Paralog	Libraries							
	POD	Root	Root2	Leaves	Nodules	Seed	Stem	Flower
GM02G47330	7.6430	6.9183	2.5093	3.2221	7.9020	4.0291	5.3964	7.1068
GM03G29440	3.5711	123.724	44.3029	6.0874	67.1994	1.5873	15.228	112.693
GM04G36410	11.8234	1.4969	0.7265	0.1815	1.1246	1.7185	5.6592	0.7955
GM05G02510	0.2077	0.4732	0.2389	0.0075	0.0455	2.2376	2.2904	0.0397
GM05G08950	9.7063	5.6654	8.8493	11.5911	0.9541	2.9815	4.482	7.2559
GM06G18890	1.8887	0.4006	0.3446	1.5637	0.4965	12.1047	0.9577	0.3194
GM09G01940	0.0408	0.0164	0	0.4497	0	0.0170	0.0559	0.0399
GM14G01430	53.0966	19.0967	30.2682	74.8153	33.0397	7.5915	8.8870	307.593
GM19G40550	5.9891	2.9200	2.2371	6.1291	2.0190	5.3064	8.1986	4.0568

CONCLUSION

In a broader overview, we demonstrated that the RFO orthologs can be split into three phylogenetic groups. In each of these groups, one paralog of soybean was found under purifying selection while the other paralog members of each group appear to be under less evolutionary constraints. These findings strongly support Ohno's gene duplication evolution model. Additionally, only soybean paralogs from one specific phylogenetic group (Group 3) were found under negative selection, unlikely to fix non-neutral mutations from the divergence of Eudicots to Monocots. We also demonstrated specific sites susceptible to mutations and under positive selection across all eight soybean RFO paralogs, however the majority of RFOs amino acid sites appear to be under restrictive evolutionary constraints. Finally, the positively selected sites of RFOs can be better explored for soybean breeding, in order to reduce the undesirable effects of raffinose family oligosaccharides on the human consumption.

ACKNOWLEDGMENTS

The financial support for this study was provided by the Minas Gerais State Foundation of Research Aid - FAPEMIG (PPM 00561–15) to LOO. LOO received a fellowship from CNPq (PQ 304153/2012-5); HVSR received a fellowship from CAPES.

REFERENCES

Akaike H (1973) Information theory and an extension of maximum likelihood principle. Second International Symposium on Information Theory, Budapest: Akademiai Kiado

Bhatt S, Katzourakis A, Pybus OG (2010) Detecting natural selection in {RNA} virus populations using sequence summary statistics. Infect Genet Evol 10:421–430. doi: http://dx.doi.org/10.1016/j.meegid.2009.06.001

Blanc G, Wolfe KH (2004) Functional Divergence of Duplicated Genes Formed

by Polyploidy during Arabidopsis Evolution. Plant Cell 16:1679–1691. doi: 10.1105/tpc.021410

Burrows M, Wheeler DJ (1994) A Block-sorting Lossless Data Compression Algorithm.

Castillo EM, De Lumen BO, Reyes PS, De Lumen HZ (1990) Raffinose synthase and galactinol synthase in developing seeds and leaves of legumes. J Agric Food Chem 38:351–355. doi: 10.1021/jf00092a003

Collakova E, Aghamirzaie D, Fang Y, et al (2013) Metabolic and Transcriptional Reprogramming in Developing Soybean (Glycine max) Embryos. Metabolites 3:347–372. doi: 10.3390/metabo3020347

Dierking EC, Bilyeu KD (2009) New sources of soybean seed meal and oil composition traits identified through TILLING. BMC Plant Biol 9:89. doi: 10.1186/1471-2229-9-89

Dierking EC, Bilyeu KD (2008) Association of a Soybean Raffinose Synthase Gene with Low Raffinose and Stachyose Seed Phenotype. 135–145.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi: 10.1093/nar/gkh340

Edger PP, Pires JC (2009) Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosom Res 17:699–717. doi: 10.1007/s10577-009-9055-9

Egert A, Keller F, Peters S (2013) Abiotic stress-induced accumulation of raffinose in Arabidopsis leaves is mediated by a single raffinose synthase (RS5, At5g40390). BMC Plant Biol 13:218. doi: 10.1186/1471-2229-13-218

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol Bioinforma 1:47–50.

Fu YX (1997) Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147:915–925.

Garrison E, Marth G (2012) Freebayes.

Goodstein DM, Shu S, Howson R, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–86. doi: 10.1093/nar/gkr944

Haldane JBS (1932) The Causes of Evolution. Princeton University Press

Hudson CM, Puckett EE, Bekaert M, et al (2011) Selection for higher gene copy number after different types of plant gene duplications. Genome Biol Evol 3:1369–1380. doi: 10.1093/gbe/evr115

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108. doi: 10.1038/nrg2689

Jiao Y, Wickett NJ, Ayyampalayam S, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100. doi: 10.1038/nature09916

Koboldt DC, Chen K, Wylie T, et al (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–5. doi: 10.1093/bioinformatics/btp373

Lam H-M, Xu X, Liu X, et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–9. doi: 10.1038/ng.715

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi: 10.1186/gb-2009-10-3-r25

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9. doi: 10.1093/bioinformatics/btp352

Li X, Zhuo J, Liu X, Wang X (2011) Expression of a GALACTINOL SYNTHASE gene is positively associated with desiccation tolerance of Brassica napus seeds during development. J Plant Physiol 168:1761–1770. doi: 10.1016/j.jplph.2011.04.006

Li Z, Baniaga AE, Sessa EB, et al (2015) Early genome duplications in conifers and other seed plants. Sci Adv 1:e1501084–e1501084. doi: 10.1126/sciadv.1501084

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452.

Maere S, De Bodt S, Raes J, et al (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454–5459. doi: 10.1073/pnas.0501102102

McCormack JE, Hird SM, Zellmer AJ, et al (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol 66:526–538.

McDonald JH, Kreitman M (1997) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654. doi: 10.1038/351652a0

McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–303. doi: 10.1101/gr.107524.110

Nylander JAA (2004) MrModeltest v2.

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York

Peterbauer T, Mucha J, Mach L, Richter A (2002) Chain Elongation of Raffinose in Pea Seeds. Isolation, characterization, and molecular cloning of a multifunctional enzyme catalyzing the synthesis of stachyose and verbascose. J Biol Chem 277:194–200. doi: 10.1074/jbc.M109734200

Pond SLK, Frost SDW (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–3. doi: 10.1093/bioinformatics/bti320

Proost S, Van Bel M, Vaneechoutte D, et al (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43:D974–81. doi: 10.1093/nar/gku986

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi:

doi:10.1038/nature08670

Shoemaker RC, Schlueter J, Doyle JJ (2006) Paleopolyploidy and gene duplication in soybean and other legumes. Curr Opin Plant Biol 9:104–9. doi: 10.1016/j.pbi.2006.01.007

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–95.

Veitia RA (2002) Exploring the etiology of haploinsufficiency. Bioessays 24:175–84. doi: 10.1002/bies.10023

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. In: 1Evolution.

Analysis of raffinose family oligosaccharide orthologs in flowering plants re-sheds light on Ohno's gene duplicate evolution model

SUPPLEMENTARY MATERIAL

Supplementary Table S1. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of raffinose family oligosaccharides (RFO) paralogs of 32 soybeans, using ortholog sequences of *A. trichopoda* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	*P*-value
	Fixed differences between species	Polymorphic sites	Fixed differences between species	Polymorphic sites		
GM02G47330	4	69	2	130	3.768	0.189
GM03G29440	2	70	6	130	0.618	0.716
GM04G36410	5	67	4	130	2.425	0.282
GM05G08950	3	67	6	127	0.948	1
GM05G02510	3	67	10	126	0.564	0.549
GM06G18890	4	67	7	126	1.075	1
GM09G01940	1	67	9	129	0.214	0.170
GM14G01430	3	67	2	126	2.821	0.348
GM19G40550	6	67	11	127	1.034	1

Supplementary Table S2. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of raffinose family oligosaccharides (RFO) paralogs of 32 soybeans, using ortholog sequences of *O. sativa* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	*P*-value
	Fixed differences between species	Polymorphic sites	Fixed differences between species	Polymorphic sites		
GM02G47330	40	140	34	197	1.655	0.053
GM03G29440	42	135	19	191	3.127	0.000*
GM04G36410	44	135	24	188	2.553	0.000*
GM05G08950	36	137	41	184	1.179	0.524
GM05G02510	31	136	32	182	1.296	0.404
GM06G18890	37	137	30	182	1.638	0.079
GM09G01940	49	134	30	183	2.231	0.002*
GM14G01430	42	136	32	191	1.843	0.019*
GM19G40550	53	134	61	183	1.187	0.442

Supplementary Table S3. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of raffinose family oligosaccharides (RFO) paralogs of 32 soybeans, using ortholog sequences of *A. thaliana* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	*P*-value
	Fixed differences between species	Polymorphic sites	Fixed differences between species	Polymorphic sites		
GM02G47330	18	163	33	293	0.980	1
GM03G29440	16	166	22	291	1.275	0.487
GM04G36410	37	161	65	284	1.004	1
GM05G08950	25	160	52	285	0.856	0.607
GM05G02510	21	160	38	280	0.967	1
GM06G18890	26	163	45	281	0.996	1
GM09G01940	36	160	58	285	1.106	0.723
GM14G01430	18	160	30	289	1.084	0.874
GM19G40550	26	159	40	280	1.145	0.681

Supplementary Table S3. McDonald and Kreitman Test (MKT) from coding-DNA sequence alignment of raffinose family oligosaccharides (RFO) paralogs of 32 soybeans, using ortholog sequences of *M. truncatula* as out-group.

Paralog	Synonymous substitutions		Nonsynonymous substitutions		NI	*P*-value
	Fixed differences between species	Polymorphic sites	Fixed differences between species	Polymorphic sites		
GM02G47330	6	86	5	176	2.456	0.190
GM03G29440	3	86	0	178	-	0.036*
GM04G36410	1	85	4	174	0.512	0.671
GM05G08950	5	84	9	172	1.138	1
GM05G02510	5	85	7	170	1.446	0.542
GM06G18890	4	85	6	171	1.341	0.735
GM09G01940	7	84	13	176	1.128	0.807
GM14G01430	8	84	6	171	2.714	0.082
GM19G40550	7	84	5	171	2.850	0.114

Supplementary Table S5. Pearson's correlation matrix of Fragments Per Kilobase of exon per Million mapped reads (FPKM) values of nine raffinose family oligosaccharides (RFO) paralogs of soybeans.

	GM14G01430	GM03G29440	GM04G36410	GM09G01940	GM05G02510	GM06G18890	GM05G08950	GM19G40550
GM02G47330	0.29976968	0.52583429	0.43529811	-0.55536573	-0.1352758	-0.19378297	-0.167524	0.06267616
GM14G01430		0.48888853	-0.16084237	-0.05513764	-0.3826251	-0.23416118	0.2901315	-0.0160093
GM03G29440			-0.34568501	-0.44809899	-0.3354472	-0.3888079	-0.07480772	-0.42030575
GM04G36410				-0.26933233	0.1749891	0.01113411	0.22888469	0.5270691
GM09G01940					-0.3001957	-0.18943083	0.12889208	-0.0631874
GM05G02510						0.64586232	-0.35279092	0.59370058
GM06G18890							-0.21512035	0.28287392
GM05G08950								0.3460034

Both mechanism and age of duplications contribute to biased gene retention patterns in plants

Both mechanism and age of duplications contribute to biased gene retention patterns in plants

Hugo V. S. Rody,[1,3] Gregory J. Baute,[2,3] Loren H. Rieseberg[2], and Luiz O. Oliveira,[1,4]

[1]Department of Biochemistry and Molecular Biology, Federal University of Viçosa, Viçosa, Minas Gerais 36570-900, Brazil; [2]Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[3]These authors contributed equally to this work.
[4]Corresponding author.

Luiz Orlando de Oliveira
Departamento de Bioquímica e Biologia Molecular
Av. P. H. Rolfs s/n
Universidade Federal de Viçosa
36570-000 Viçosa (MG), Brasil
Tel.: 55 31 3899 2964
Email: lorlando@ufv.br

Running Title: Low turnover rates bias duplicate gene retention

Key words: biased gene retention, duplicate genes, genomic balance, molecular evolution, polyploidy, tandem duplication, turnover rate, whole-genome duplication

ABSTRACT

All extant seed plants are successful paleopolyploids, whose genomes carry duplicate genes that have survived repeated episodes of diploidization. However, the survival of gene duplicates is biased with respect to gene function and mechanism of duplication. Transcription factors, in particular, appear to be preferentially retained if duplicated by whole-genome duplications (WGDs), but disproportionately lost when duplicated by tandem events. An explanation for this pattern is provided by the Gene Balance Hypothesis (GBH), which posits that duplicates of highly connected genes are retained following WGDs to maintain optimal stoichiometry among gene products; but such connected gene duplicates are disfavored following tandem duplications. Here, we used genomic data from 25 taxonomically diverse plant species to investigate the roles of duplication mechanism, gene function, and age of duplication in the retention of duplicate genes. Enrichment analyses were conducted to identify Gene Ontology (GO) functional categories that were overrepresented in either WGD or tandem duplications, or across ranges of divergence times. Tandem paralogs were much younger, on average, than WGD paralogs and the most frequently overrepresented GO categories were not shared between tandem and WGD paralogs. Transcription factors were overrepresented among ancient paralogs regardless of mechanism of origin or presence of a WGD. Also, in many cases, there was no bias toward transcription factor retention following recent WGDs. These observations can be reconciled with the GBH if selection for optimal stoichiometry among gene products is strongest following the earliest polyploidization events and becomes increasingly relaxed as gene families expand.

INTRODUCTION

Gene duplication has long been viewed as a key driver of biological complexity in Eukaryotes (Ohno 1970; Wendel 2000; Kondrashov et al. 2002; Zhang 2003). Duplicate genes mainly arise via small-scale tandem or segmental duplication events or via large-scale whole genome duplications (WGDs). The latter are especially common in plants (Wood et al. 2009; Mayrose et al. 2011). Indeed, comparative genomic studies indicate that all extant seed and flowering plants have experienced one or more WGDs in their evolutionary history (Blanc and Wolfe 2004b; Shoemaker et al. 2006; Jaillon et al. 2007; Barker et al. 2008; Jiao et al. 2011; Li et al. 2015).

Following gene duplication (whether via tandem, segmental or WGD events), most duplicate copies become pseudogenes (i.e. lose their function) or are lost entirely due to deletions. This is expected because of relaxed purifying selection due to functional redundancy. Large-scale deletions are especially common following WGDs, as the neopolyploid returns back to its ancestral diploid condition, a process referred to as diploidization. Nevertheless, some gene duplicates are retained, and these surviving duplicates appear to contribute importantly to the evolution of biological complexity and phenotypic novelty, in part because such genes are less constrained evolutionarily than are single copy genes (Lynch and Conery 2000; Edger and Pires 2009; Edger et al. 2015).

Several models have been put forward to explain how duplicate genes avoid pseudogenization, as well as to account for why some duplicate genes are retained and others are not (reviewed in Conant et al. 2014). These include (1) neofunctionalization, in which one of the duplicates (i.e. paralogs) acquires a new function; (2) subfunctionalization, in which ancestral function is partitioned among paralogs (Ohno 1970); (3) relative dosage, in which duplicate genes are retained (or lost) to avoid dosage imbalances (Birchler et al. 2001; Freeling 2009); and (4) absolute dosage, in which the fixation of duplicate genes is due to selection favoring an increase in gene dosage (Kondrashov and Koonin 2004) or metabolic flux (Hudson et al. 2011).

In this paper, we focus on the predictions of the relative dosage model, also known as the Gene Balance Hypothesis (GBH) (Birchler et al. 2001; Veitia

2002; Papp et al. 2003; Freeling 2009), as this hypothesis has garnered the most support from real data (Blanc and Wolfe 2004a; Maere et al. 2005; Tian et al. 2005; Rodgers-Melnick et al. 2012). According to the GBH, genes with a large number of interactions (i.e., “connected genes”) should be retained disproportionately following WGD events thereby maintaining optimal stoichiometry among their products; when a WGD event occurs, all genes are duplicated simultaneously and so relative gene dosage should not change. In small-scale duplications (e.g., tandem events), the increased dosage of a single gene belonging to a gene complex may result in decreased fitness, or even in lethality. Thus, connected genes are expected to be differentially lost following small-scale duplications. Conversely, genes that work alone or have few interactions, such as those involved in disease resistance, are more likely to be retained following tandem duplications.

Patterns of gene retention in *Arabidopsis thaliana* are largely consistent with GBH predictions. For example, highly connected genes such as transcription factors have been preferentially retained after WGDs in *A. thaliana*, but disproportionately lost following small-scale duplications (Blanc and Wolfe 2004a; Maere et al. 2005). Similar findings have been reported for poplar (Rodgers-Melnick et al. 2012) and rice (Tian et al. 2005). In contrast, paleologs (paralogs arising from WGD events) in the Compositae family are reported to be enriched for genes annotated to structural components or cellular organization gene ontology (GO) categories, while genes involved with transcription appear to be significantly under-represented (Barker et al. 2008). And in *A. thaliana* and *Sorghum bicolor*, both WGD and tandem mechanisms of duplication are associated with paralogs involved in high metabolic flux networks (Hudson et al. 2011), an observation best predicted by the absolute dosage model.

In addition to mechanism of duplication, the fate of paralogs may be influenced by genetic background, various environmental factors, epigenetic effects, genetic drift, and the mechanism of gene dosage-compensation (Crow and Wagner 2006; Edger and Pires 2009; Hudson et al. 2011). Another potential issue concerns the faster turnover rates of tandem paralogs relative to those originating via WGDs (Lynch and Conery 2000; Blanc and Wolfe 2004b; Hanada et al. 2008; Wang 2013). As a consequence, the sampling of tandem paralogs is biased towards young gene duplicates whereas that of WGD

paralogs is skewed towards old duplications. As far as we are aware, this bias has not previously been accounted for when inferring patterns of duplicate gene retention. Nor has there been a comprehensive analysis of retention patterns across the plant kingdom.

Here we investigate the impact of duplication mechanism, gene function, and age of duplication in the retention of duplicate genes. Our analyses consider both WGD and tandem duplications, as these are the two most frequently invoked mechanisms to explain how paralogous gene pairs are generated in plant genomes (Kondrashov et al. 2002; Blanc and Wolfe 2004a; Maere et al. 2005; Thomas et al. 2006). We have targeted 25 plant species with fully sequenced genomes that include the basal land plants, *Physcomitrella* and *Selaginella*, the basal flowering plant *Amborella*, and as well as 14 flowering plant orders. This diverse array of taxa enables comparisons of taxa with highly contrasting histories of polyploidy, including at least one species with no known WGD in its evolutionary history (*Selaginella*). This is critical, because it allows us to control for potential biases caused by unequal duplicate gene turnover rates. Our focus is on genes annotated as transcription factors, since differential retention of duplicated transcription factors provides the main support for the GBH. We specifically address the following questions: (1) Is the turnover rate of WGD paralogs persistently lower than that of tandem paralogs? (2) Which functional gene categories are consistently overrepresented among WGD and/or tandem paralogs? (3) Does variation in duplicate gene retention depend significantly upon the age of WGD paralogs? (4) To what extent do our results support for the Gene Balance Hypothesis?

RESULTS

Origin and turnover rate of paralogs

For each of the 25 study species, we calculated *Ks* time divergence between pairs of paralogs and used a synteny-based approach to categorize members of all gene families as derived from WGD or tandem duplications. Gene families whose origins were uncertain based on available data were classified as “undefined”. Across the 25 target genomes, the majority of

paralogs detected had *Ks* ≤ 2 (Table 1) including 79% of paralogs in *A. thaliana*, 86% in *Glycine max,* and 92% in *Malus domestica*. Paralogs with *Ks* > 2 were excluded from our analyses due to concerns that *Ks* saturation could impair reliable inferences (Vanneste et al. 2013). Most species displayed clear prominent peaks in their *Ks* age histograms, which is illustrated by histograms for five species with contrasting histories of polyploidy (Figure 1; histograms for the remaining 20 species are depicted in Supplemental Figure 1). In the K-S goodness of fit test, all histograms for all species except *Carica* deviated significantly ($P < 0.05$) from the null model of constant duplicate gene birth and death (Supplemental Table 1). SiZer maps identified a significantly increasing gradient in the *Ks* age histograms of WGD-derived paralogs of most species, which provides support for polyploid signals being well distinguished from background duplications.

The *Ks* age histograms of WGD-derived paralogs (Figure 1A, depicted in black) were clearly distinct from those of the tandem-derived paralogs (Figure 1B, depicted in gray). While tandem histograms exhibited a descending slope (similar to a half-parabola) for most of the species, WGD-derived paralog histogams had peaks that overlaped with peaks from histograms of all paralogs (Figure 1A, depicted in brown). SiZer maps also confirmed the presence of peaks for WGD-derived paralogs histograms (Figure 1D).

Because of our focus on transcription factor paralogs, their *Ks* age histograms are shown (Figure 1 and Supplemental Figure 1; depicted in yellow) along with the *Ks* histograms of WGD- and tandem-derived paralogs. The SiZer maps (Figure 1D and E) showed increasing gradients for transcription factor paralogs that overlapped with the slopes of WGD-derived paralogs.

Table 1. Distribution of paralogous gene pairs for 25 plant species targeted by this study

Specie	Chr	Initial PCG	Duplicates	Number of duplicates by duplication type			Number of duplicates by Ks range				
				WGD	Tandem	Undefined	0 < Ks ≤ 0.5	0.5 < Ks ≤ 1	1 < Ks ≤ 1.5	1.5 < Ks ≤ 2	Ks > 2
Arabidopsis lyrata	16	32670	6378	3442	1816	1120	2251	2216	966	228	2266
Arabidopsis thaliana	10	33602	6194	2740	1232	2222	1657	2407	1183	222	1664
Amborella trichopoda	26	26460	3322	15	998	2309	1861	427	402	137	1542
Brachypodium distachyon	10	26678	3573	1025	1768	780	981	1024	835	175	2967
Carica papaya	18	28072	1915	24	455	1436	603	210	402	126	1588
Citrullus lanatus	22	23438	2807	385	1015	1407	691	435	751	227	2493
Eucalyptus grandis	22	36449	11120	390	6424	4306	8106	925	1029	240	2661
Fragaria vesca	14	34809	3974	1021	1606	1347	1500	979	684	184	2095
Glycine max	40	46509	15242	9721	2087	3434	11697	1961	790	185	2388
Helianthus annuus	34	97436*	38196	998	1629	35569	25185	9991	2502	37	1074
Lotus japonicus	12	26818	2682	184	627	1871	1159	774	415	51	1161
Malus domestica	34	63515	15551	2761	1308	11482	13084	1258	683	107	1286
Manihot esculenta	36	30800	7134	2530	703	3901	4915	837	716	110	1174
Medicago truncatula	16	57587	5098	1083	2419	1596	2902	1262	543	115	1880
Oryza sativa ssp. indica	24	48788	8349	1957	2869	3523	3361	2169	1665	317	4182
Oryza sativa ssp. japonica	24	59430	5559	1482	2173	1904	1928	1584	1233	183	3048
Physcomitrella patens	54	36137	3769	306	202	3261	637	1848	883	99	830
Populus trichocarpa	36	41521	9721	5609	1988	2124	7572	738	704	147	1924
Ricinus communis	20	31221	2558	155	614	1789	628	435	683	176	2096
Sorghum bicolor	20	34686	4267	1048	1698	1521	1468	1061	993	186	3046
Solanum lycopersicum	24	34432	7100	1234	2561	3305	3184	2287	872	209	3061
Selaginella moellendorffii	16-27	22285	1885	351	608	926	1457	129	102	66	801
Theobroma cacao	20	46269	3488	722	1553	1213	1199	601	822	201	2228
Vitis vinifera	38	26644	4536	528	1935	2073	1918	852	1042	128	1602
Zea mays	20	39597	6336	590	1396	4350	3792	1095	813	153	2673

Chr, Number of Chromosomes; Initial PCG, Initial number of Protein-coding gene sequences
*Including alternative transcrips

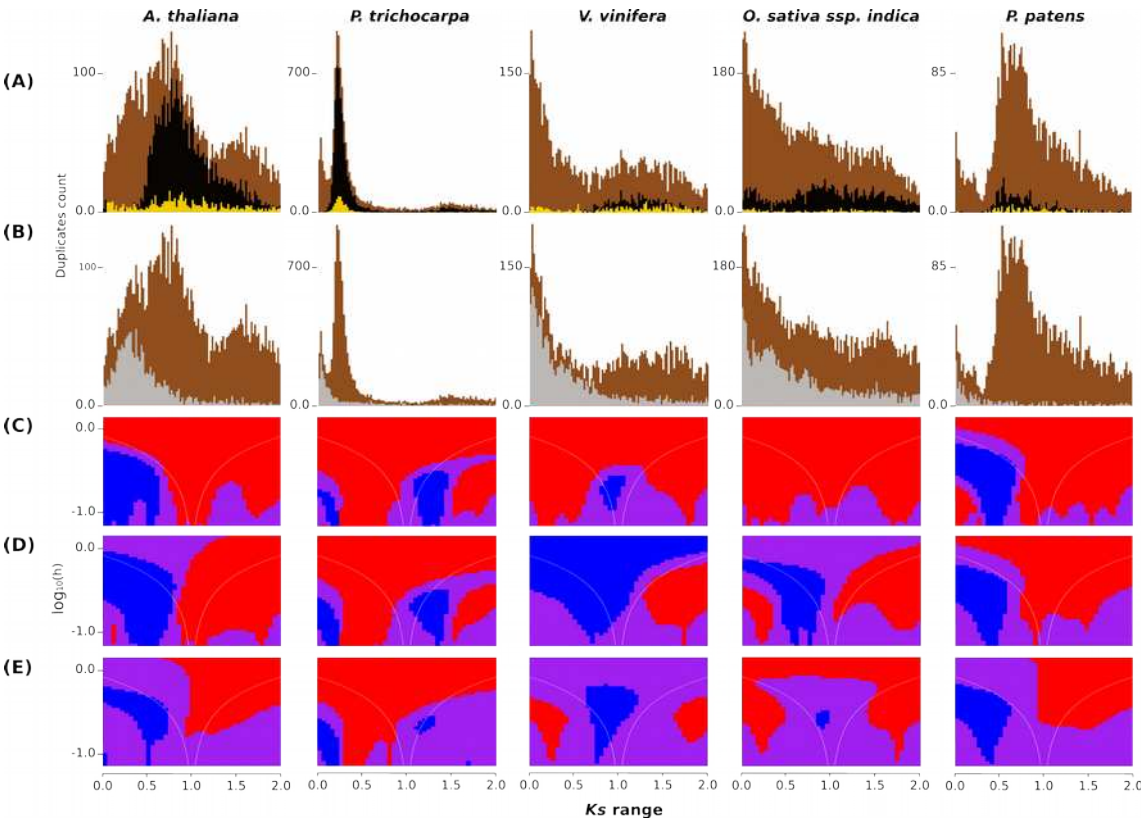


Figure 1. *Ks* age distributions (A and B) and SiZer maps (C to E) of five plant species. (A) Brown bars, all paralogs (background); black bars, WGD-derived paralogs predicted by DAGchainer; yellow bars, paralogs annotated as transcription factor activity (GO:0003700). (B) Brown bars, background; gray bars, tandem-derived paralogs predicted by DAGchainer. SiZer maps for (C) All paralogs; (D) WGD-derived paralogs; (E) Transcription factor paralogs.

Biased retention of paralogs after large- and small-scale duplications

To assess the universality of the GBH across land plants, we submitted the predicted WGD- and tandem-derived paralogs from the 25 target genomes to enrichment analyses and identified the most strongly overrepresented GO functional categories. We found that WGD- and tandem-derived paralogs did not share the top 10 most frequently overrepresented GO categories (Figure 2A and C). While the most overrepresented categories of WGD-derived paralogs fell under macromolecular complexes (GO:0032991), internal to cell (GO:0005622), and cytoplasm (GO:0005737) functional GO categories; those of tandem-derived paralogs grouped into programmed cell death

(GO:0012501), defense response (GO:0006952), and apoptotic process (GO:0006915) GO categories.

In six species, WGD-derived paralogs were not enriched for the overrepresented GO categories found in the remaining plant species. Five of them — *Cariaca*, *Ricinus, Populus*, *Selaginella*, and *Physcomitrella* — have few WGD-derived paralogs predicted by DAGchainer (Table 1), consistent with possible under-estimation or misidentification of WGD-derived paralogs in these species (see discussion below). For another five taxa — *A. thaliana*, *Medicago,* both *Oryza* subspecies*,* and *Populus* — WGD-derived transcription factor paralogs were overrepresented (Figure 2B). Surprisingly, WGD-derived transcription factor paralogs were not significantly overrepresented in *Arabidopsis lyrata*, which shares the same WGD events as *A. thaliana*, although there was a trend in the expected direction.

Unexpectedly, transcription factor activity (GO:0003700) WGD-derived paralogs were not significantly overrepresented in 20 plant species, ten of which exhibit evidence of recent WGDs in their evolutionary history, with a significantly increasing gradient in SiZer (Figure 1, Supplemental Figure 1) within Ks range < 1 (and consistent with previous reports — see below). Finally, results from our analyses of tandem duplications showed tandem-derived transcription factor paralogs were significantly underrepresented across the 25 focal genomes.

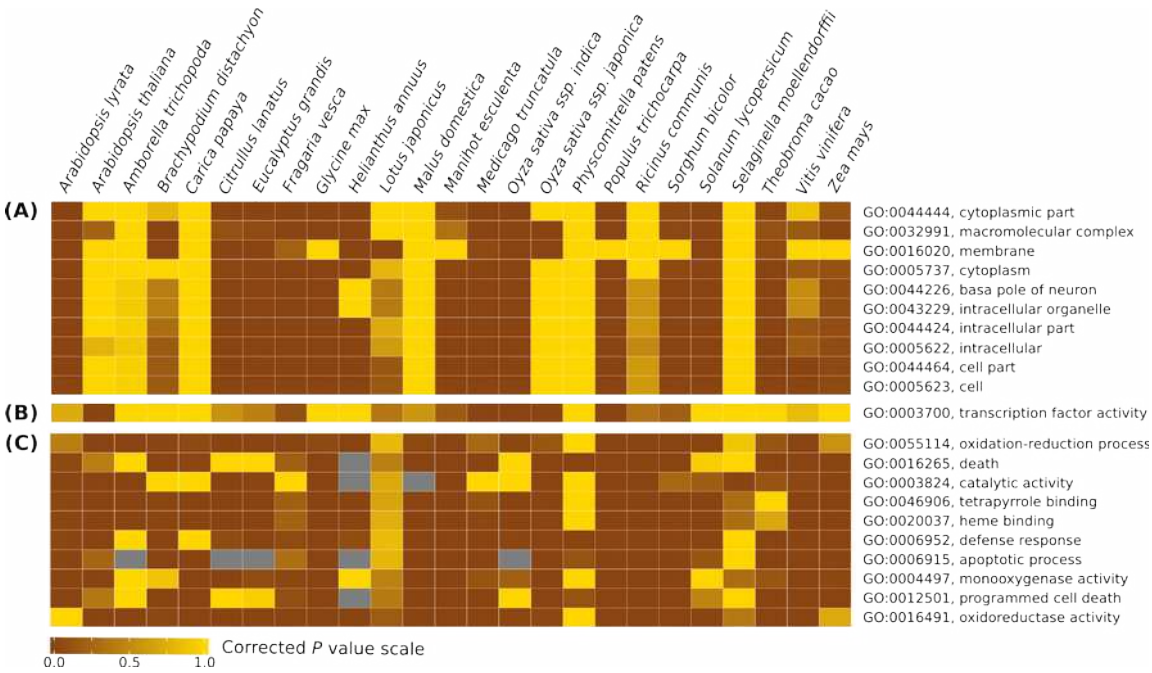


Figure 2. Heat maps of GO categories across 25 plant species. (A) The 10 most frequent GO categories overrepresented among WGD-derived paralogs. (B) Transcription factor activity category (GO:0003700) enrichment analysis for WGD-derived paralogs. (C) The 10 most frequent GO categories overrepresented among tandem-derived paralogs. Color gradient represents the Corrected *P* value calculated by the ErmineJ software: brown colors, significant over-representation (*P<0.05*); yellow colors, reduced or non-significant enrichment; and gray color, no enrichment.

Biased retention toward ancient transcription factors

We analyzed the biased retention of transcription factor paralogs based on *Ks* time divergence as opposed to mechanism of duplication. This was accomplished by mapping known WGD events onto a phylogeny for the 25 species targeted by this study (Figure 3, Supplemental Table 2).

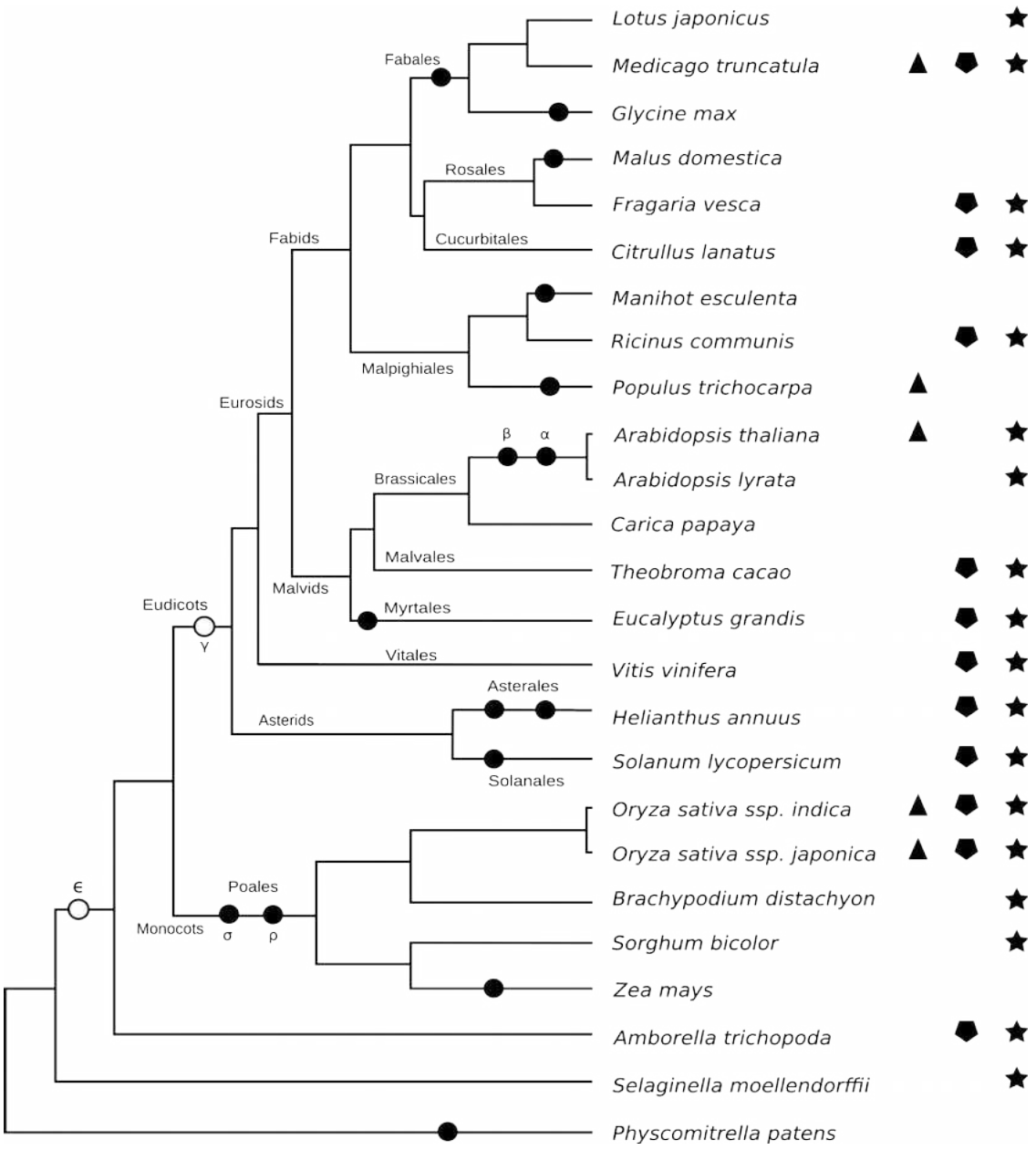


Figure 3. Phylogenetic distribution of transcription factor retention biases among 25 plant species. The phylogenetic tree was adapted from PLAZA 3.0. Symbol code: Black circles on the tree branches, all known WGD events we also identified in this study; Open circles, suggested ancient WGD events we did not examine; triangles, species with WGD-derived transcription factor paralogs significantly overrepresented; pentagons and stars, species with transcription factor paralogs significantly overrepresented in range $1.5 < Ks \leq 2$ and range $1 < Ks \leq 2$, respectively.

In general, old (*Ks* > 1) transcription factor (GO:003700) paralogs tend to be overrepresented in the genomes investigated here (Figure 3). Twelve of the 25 focal species exhibited significant enrichment at *Ks* range > 1.5 (Figure 3, pentagons), but no such retention bias at lower *Ks* ranges (≤ 1.5). When we compared transcription factor paralog enrichment at *Ks* > 1.0 versus < 1.0, 18 species showed significant enrichment for the older transcription factor paralogs (Figure 3, stars). For four of these, *A. thaliana*, *Medicago*, and the two *Oryza* subspecies, the overrepresented transcription factors originated from WGD events (Figure 2B). However, for the remaining 14 species, the ancient paralogs are not obviously associated with a WGD event. Although *A. thaliana, Oryza sativa ssp. indica,* and *Solanum* exhibited significant signals of polyploidy in the *Ks* range < 1 (Figure 1; Supplemental Figure 1), their transcription factor paralogs were only significantly overrepresented in the *Ks* range > 1 (Figure 3F).

In genomes of only four taxa (*Carica*, *Malus*, *Manihot,* and *Populus*) were recent transcription factor paralogs overrepresented, and only for *Populus* were WGD-derived transcription factor paralogs significantly overrepresented (Figure 2B).

In addition to analyzing the retention of transcription factor paralogs, we submitted our data to enrichment analysis aiming to find additional GO categories that could have experienced biased retention patterns. A number of GO categories, including those involved in transcription, regulation, transport, and response to stimulus were frequently overrepresented among ancient paralogs (*Ks* > 1) (Supplemental Table 3). While three of these functional GO categories — cell periphery (GO:0071944), plasma membrane (GO:0005886), and response to abiotic stimulus (GO:0009628) — were overrepresented among WGD-derived paralogs; two categories — response to stimulus (GO:0050896) and catalytic activity (GO:0003824) — were overrepresented among tandem-derived paralogs.

DISCUSSION

Our synteny-based approach identified pairs of WGD-derived genes similar to those that have been reported in previous studies. In *A. thaliana*, for example, circa 80% of the 2740 duplicate gene pairs we classified as WGD-derived are in common with the list of polyploidy-derived paralogs published by (Blanc and Wolfe 2004b). Differences among studies may be due to new gene annotation tools that recently became available. In some instances, the number of paralogs predicted as having their origin in WGD events can be underestimated due to widespread genomic changes (e.g., gene loss and/or chromosomal rearrangements) after polyploidization events (Freeling 2009). Such processes are particularly problematic for ancient polyploidization events, which may explain the low number of WGD paralogs we predicted in the basal plants, *Amborella* and *Physcomitrella*, as well as for *Lotus*, *Carica*, and *Ricinus*. On the other hand, our approach indicates the presence of a small number WGD-derived paralogs in *Selaginella,* which is not thought to have a WGD in its evolutionary history. This result could be evidence that *Selaginella* in fact is an ancient polyploid or that some unknown fraction of WGD derived paralogs are false positives.

Tandem paralogs were similarly identified based on the genomic coordinates of genes. In *Eucalyptus*, 32% of its 36,449 protein-coding genes originated via tandem events, which is the largest proportion of tandem-derived paralogs amongst the 25 plant species we investigated. *Physcomitrella* exhibited the smallest proportion (~1%) of tandem-derived paralogs. These findings are very similar to those previously reported for *Eucalyptus* (Myburg et al. 2014) and *Physcomitrella* (Rensing et al. 2008), respectively.

The peaks we identified in the *Ks* age histograms likely resulted from WGD events. Previous studies have also identified these WGD events using data that span across several families (Vanneste et al. 2014), or from a given plant species (e.g., Jaillon et al. 2007; Barker et al. 2008; Myburg et al. 2014). In the *Ks* histogram of *A. thaliana*, for example, there were two prominent peaks (Figure 1), which coincided with the α and β polyploid events reported by early investigations (Simillion et al. 2002; Bowers et al. 2003; Vanneste et al. 2014). In our analysis, the tail of the most recent duplication masked the second peak;

thus, a single, significantly increasing slope was identified by SiZer. In *A. lyrata,* SiZer identified two significant peaks as expected given the recent history of polyploidy in *Arabidopsis* (Bowers et al. 2003).

Differences in the *Ks* age histograms from WGD- and tandem-derived paralogs indicates that the turnover rate of tandem paralogs is faster than that of WGD paralogs, as previously suggested by others (Lynch and Conery 2000; Blanc and Wolfe 2004b; Rensing et al. 2008; Wang 2013). The pattern we uncovered suggests lower turnover rates of transcription factor paralogs than those observed for tandem paralogs. Furthermore, it appears that the origin and biased retention of transcription factor paralogs are not restricted to large-scale duplication events.

Consistent with the expectations of the GBH, WGD- and tandem-derived paralogs did not share the top 10 most frequently overrepresented GO categories. Six species — *Malus, Cariaca*, *Ricinus, Populus*, *Selaginella*, and *Physcomitrella* — were exception and did not share the most frequent GO categories, which is consistent with the possible under-estimation or misidentification of WGD-derived paralogs in these species. For *Malus*, for example, the GO categories that were overrepresented include: plasma membrane (GO:0005886), response to abiotic stimulus (GO:0009628), response to biotic stimulus (GO:0009607), and response to endogenous stimulus (GO:0009719). Analyses of an EST library of *Malus domestica* also found that these categories were overrepresented (Sanzol 2010). Consistent with the GBH, we did not found tandem-derived transcription factor paralogs overrepresented in any of 25 focal genomes.

Other findings were inconsistent with the predictions of the GBH. Unexpectedly, transcription factor activity (GO:0003700) WGD-derived paralogs were only significantly overrepresented in five plant species — *A. thaliana*, *Medicago,* the two *Oryza* subspecies*,* and *Populus*. Ten of the 20 remaining study species exhibited evidence for recent WGDs. Other studies have also reported downward biased transcription factor paralogs following WGD events. In Compositae paleologs, for example, genes involved with structural components or cellular organization were significantly overrepresented; whereas transcription factors were significantly underrepresented (Barker et al. 2008). These authors argued that patterns of intrinsic selection on different

gene categories may vary across higher taxonomic categories. Hudson et al. (2011) suggested that the fate of paralogs originated by either WGD or small-scale events would depend on intrinsic properties, such as gene function and the environment in which the new polyploid was born.

Regardless the mechanism of duplication, we showed that ancient paralogs of transcription factors were preferentially retained over paralogs of more recent origin. In agreement to our findings, a previous study in *A. thaliana* reported that genes involved in transcriptional regulation showed greater retention after the later (β) genome duplication than after the youngest (α) duplication (Maere et al. 2005). Again, our results indicate that plant species with very different histories of polyploidy, such as *A. thaliana* with two recent WGD events (Simillion et al. 2002) and *Vitis* with no known recent WGD events (Jaillon et al. 2007), share this pattern of biased retention towards ancient transcription factor paralogs.

Although transcription factor paralogs with recent origin were over-represented in four species (*Carica*, *Malus*, *Manihot,* and *Populus*), we could only clearly determine that those of *Populus* were WGD-derived paralogs. The over-representation of young (*Ks* < 0.5) transcription factor paralogs in *Carica* is intriguing, given that no WGD events likely took place in its recent evolutionary history (Ming et al. 2008) and that DAGchainer only predicted tandem-derived transcription factor paralogs for *Carica* within the *Ks* range **≤** 1.0. Given that *Carica* lacks recent WGD events (Ming et al. 2008) and we did not identified transcription factors paralogs originated from WGD events within *Ks* < 1, the many transcription factor paralogs of *Carica* appear to derive from small-scale duplications within its genome.

Our analyses imply that both the fixation and retention of duplicated genes are context-dependent events. Thus, while the mechanism of duplication is clearly important, so are the characteristics of the particular lineage in which the duplication arises, as well as timing of duplication.

Although our results show that many transcription factor paralogs do indeed derive from large-scale duplication events, this is not conclusive evidence for the GBH. Observations seemingly inconsistent with the GBH include, for example, the preferential retention of transcription factor paralogs in taxa with no apparent history of polyploidy or following tandem duplications in

Carica, as well as the absence of such retention biases following some recent WGDs (e.g. *Glycine, Helianthus,* and *Zea*).

Nonetheless, the most important observation in this paper — the strong bias toward ancient transcription factor duplicates seen in most plant genomes — can be interpreted in a manner consistent with the GBH. Possibly, all plant lineages are the product of multiple ancient WGDs, the earliest of which are no longer detectable. Under the GBH, the duplicates from the first polyploidization would be most likely to be retained to maintain optimal stoichiometry among gene products. The number of paralogs is expected to grow rapidly with each polyploidization event. With so many paralogs, changes in the amount of the gene product might be tolerated and a copy of the gene can be lost or diverge. This could lead to the pattern we see — biased retention toward ancient transcription factor paralogs — and also might account for the weaker signal we see among recent transcription factor paralogs. It even could account, in part, for the greater tolerance of recent tandem transcription paralogs seen in *Carica*.

METHODS

Data collection and selection of paralogs

Full genome annotations, protein-gene codes, DNA sequences, gene families, and Gene Ontology (GO) annotations from the 25 focal species were retrieved from PLAZA 2.5 and 3.0 Dicots (Proost et al. 2015), with the exception of sunflower (*Helianthus annuus*), as detailed in Supplemental Table 4. Protein-gene code files with alternative transcripts removed were used to identify paralogous gene pairs using BLASTp all-against-all, with an e-value cutoff of e^{-20}, with a minimum 50% identity, alignment length > 300 bp, number of mismatches < 550, and number of gap opens < 30. Self hits were removed and only paralogous gene pairs with both copies belonging to the same gene family were maintained for further analysis. For the selection of paralog pairs for *H. annuus*, CDS sequences and BLASTn all-against-all were used based on HA412.v1.1.bronze version of the genome (http://www.sunflowergenome.org/).

Determining paralog duplication mechanism

The DAGchainer software package (Haas et al. 2004) was used to predict the mechanism of by which paralogs originated based on their genomic coordinates. WGD-derived paralog pairs were predicted by running DAGchainer to find syntenic/collinear regions among chromosomes, in the same species, using default parameters and ignoring tandem duplication alignments (-s and -I options). Tandem-derived paralogs were predicted by using the accessory segmental duplication tool, also made available by DAGchainer, to find collinear sets of homologous genes, with the 'max intervening genes value' set to 10. All the other paralog pairs, not predicted as WGD or tandem-derived, were marked as undefined (UD), as these paralogs may have been originated by either large- or small-scale duplications.

Age of duplication events

We calculated relative divergence times for each paralog pair in terms of synonymous substitutions per synonymous site (*Ks*). First, we aligned the nucleotide sequences of gene pairs using TranslatorX (Abascal et al. 2010), based on protein alignments performed by MUSCLE v3.8.31 (Edgar 2004). Divergence times (*Ks*) were calculated following Yang and Nielsen (2000), implemented with the *yn00* software from the PAML v4.1 package (Yang 2007). This method assumes the F3x4 codon frequency model and accounts for transition/transversion rate bias and codon usage bias, which is an approximation of the maximum likelihood method recommended for pairwise comparisons in the manual of PAML. Because of issues associated with *Ks* saturation and stochasticity (Vanneste et al. 2013), only paralogs with $Ks \leq 2$ and Standard Error (SE) < 0.5 were used in further analyses.

Custom python scripts were used to parse the BLAST all-against-all output in order to identify the closest paralog gene pairs. First, self hits were removed. Then, paralogs were organized into a single gene list and then used to select the corresponding paralog pair(s) for each of these genes based on the following three rules: (I) if a single gene was predicted as WGD-derived by DAGchainer, keep the duplicate pair with the lowest *Ks* value, while still allowing

pairing with tandem-derived or undefined genes; (II) if not predicted as WGD-derived, but predicted as tandem-derived, keep the gene pair with the lowest *Ks* value; and (III) if the single gene was not predicted as WGD- or tandem-derived, keep the undefined paralog pair with the lowest *Ks* value.

GO annotation and over-representation analysis

Functional Gene ontology GO terms (categories) were determined for each gene and paralog pair and then evaluated for enrichment by the ErmineJ v3.0.2 software (Gillis et al. 2010). All the three GO domains (Biological Process, Molecular Function and Cellular Component) were included in the Over-Representation Analysis (ORA), with a minimum gene set size equal to 10 and the Best Scoring gene replicate treatment. Eight different groups of paralogs were analyzed: WGD-derived, tandem-derived, and paralogs representing the following *Ks* ranges: (A) $0 < Ks \leq 0.5$, (B) $0.5 < Ks \leq 1$, (C) $1 < Ks \leq 1.5$, (D) $1.5 < Ks \leq 2$, (E) $0 < Ks \leq 1$ and (F) $1 < Ks \leq 2$. The GO categories were considered overrepresented if Corrected $P < 0.05$, as calculated by the ErmineJ software.

Statistics

K-S goodness of fit test

The Kolmogorov-Smirnov test (Cui et al. 2006) was used to evaluate if the age distribution (*Ks*) of all duplicates (background) deviated significantly ($P < 0.05$) from a simulated null hypothesis of constant duplicate gene birth and death.

SiZer maps: identifying significant peaks in *Ks* histograms

Significant peaks in the *Ks* histograms were found by SiZer (Chaudhuri and Marron 2000) implemented on R software, with the following command line: SiZer.1 <- SiZer(x, y, h=c(.05,5), degree=1, derv=1). A SiZer map is a way of examining when the p-th derivative of a scatterplot-smoother is significantly

negative, possibly zero or significantly positive across a range of smoothing bandwidths. In a SiZer map, blue indicates a significantly increasing gradient, red is a significantly decreasing gradient, purple is a non-significant gradient and gray indicates that data are too sparse for reliable estimation.

DATA ACCESS

Data generated in this study are available in as a single file at ftp://ftp.ufv.br/dbb/geneduplication/.

ACKNOWLEDGMENTS

The financial support for this study was provided by the Minas Gerais State Foundation of Research Aid - FAPEMIG (PPM 00561–15) to LOO. LOO received a fellowship from CNPq (PQ 304153/2012-5); HVSR received a fellowship from the Brazilian program for research incentives "*Science without borders*". LHR was supported by grants from Genome Canada and Genome BC.

REFERENCES

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**: W7–13.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* **25**: 2445–2455.

Birchler J a, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol* **234**: 275–288.

Blanc G, Wolfe KH. 2004a. Functional Divergence of Duplicated Genes Formed

by Polyploidy during Arabidopsis Evolution. *Plant Cell* **16**: 1679–1691.

Blanc G, Wolfe KH. 2004b. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* **16**: 1667–1678.

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

Chaudhuri P, Marron JS. 2000. Scale space view of curve estimation. *Ann Stat* **28**: 408–428.

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.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**: 738–749.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

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 U S A* **112**: 8362–6.

Edger PP, Pires JC. 2009. Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosom Res* **17**: 699–717.

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.

Gillis J, Mistry M, Pavlidis P. 2010. Gene function analysis in complex data sets using ErmineJ. *Nat Protoc* **5**: 1148–1159.

Haas BJ, Delcher AL, Wortman JR, Salzberg SL. 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**: 3643–3646.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiol* **148**: 993–1003.

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.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Jiao Y, Wicket N, Ayyampalayam S, Chanderbali A. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kondrashov FA, Koonin E V. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20**: 287–90.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin E V. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3**: research0008.l0008.9.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Sci Adv* **1**: e1501084–e1501084.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science (80-)* **290**: 1151–1155.

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 U S A* **102**: 5454–5459.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science (80-)* **333**: 1257.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly B V, Lewis KLT, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**: 991–996.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. 2014. The genome of Eucalyptus grandis. *Nature* **510**: 356–362.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. 2015. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res* **43**: D974–81.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y, et al. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science (80-)* **319**: 64–69.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, Difazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Res* **22**: 95–105.

Sanzol J. 2010. Dating and functional characterization of duplicated genes in the apple (Malus domestica Borkh.) by analyzing EST data. *BMC Plant Biol* **10**: 87.

Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Curr Opin Plant Biol* **9**: 104–9.

Simillion C, Vandepoele K, Montagu MCE Van, Zabeau M, Peer Y Van De. 2002. The hidden duplication past of Arabidopsis thaliana. *PNAS* **99**:

13627–13632.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16**: 934–946.

Tian C-G, Xiong Y-Q, Liu T-Y, Sun S-H, Chen L-B, Chen M-S. 2005. Evidence for an ancient whole-genome duplication event in rice and other cereals. *Yi Chuan Xue Bao* **32**: 519–527.

Vanneste K, Baele G, Maere S, Van De Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res* **24**: 1334–1347.

Vanneste K, Van De Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Mol Biol Evol* **30**: 177–190.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* **24**: 175–84.

Wang Y. 2013. Locally duplicated ohnologs evolve faster than nonlocally duplicated ohnologs in Arabidopsis and rice. *Genome Biol Evol* **5**: 362–9.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol Biol* **42**: 225–249.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci U S A* **106**: 13875–9.

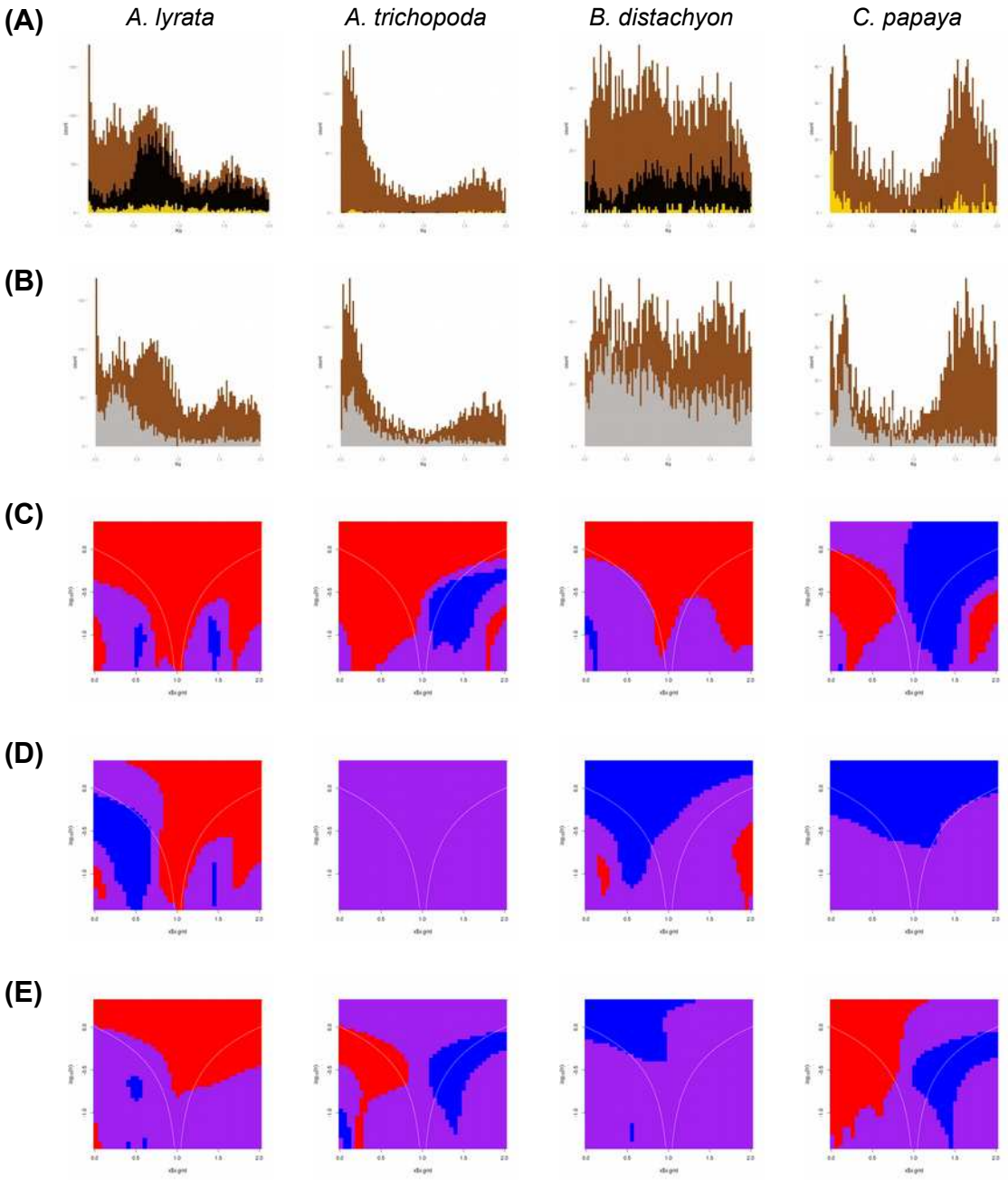
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

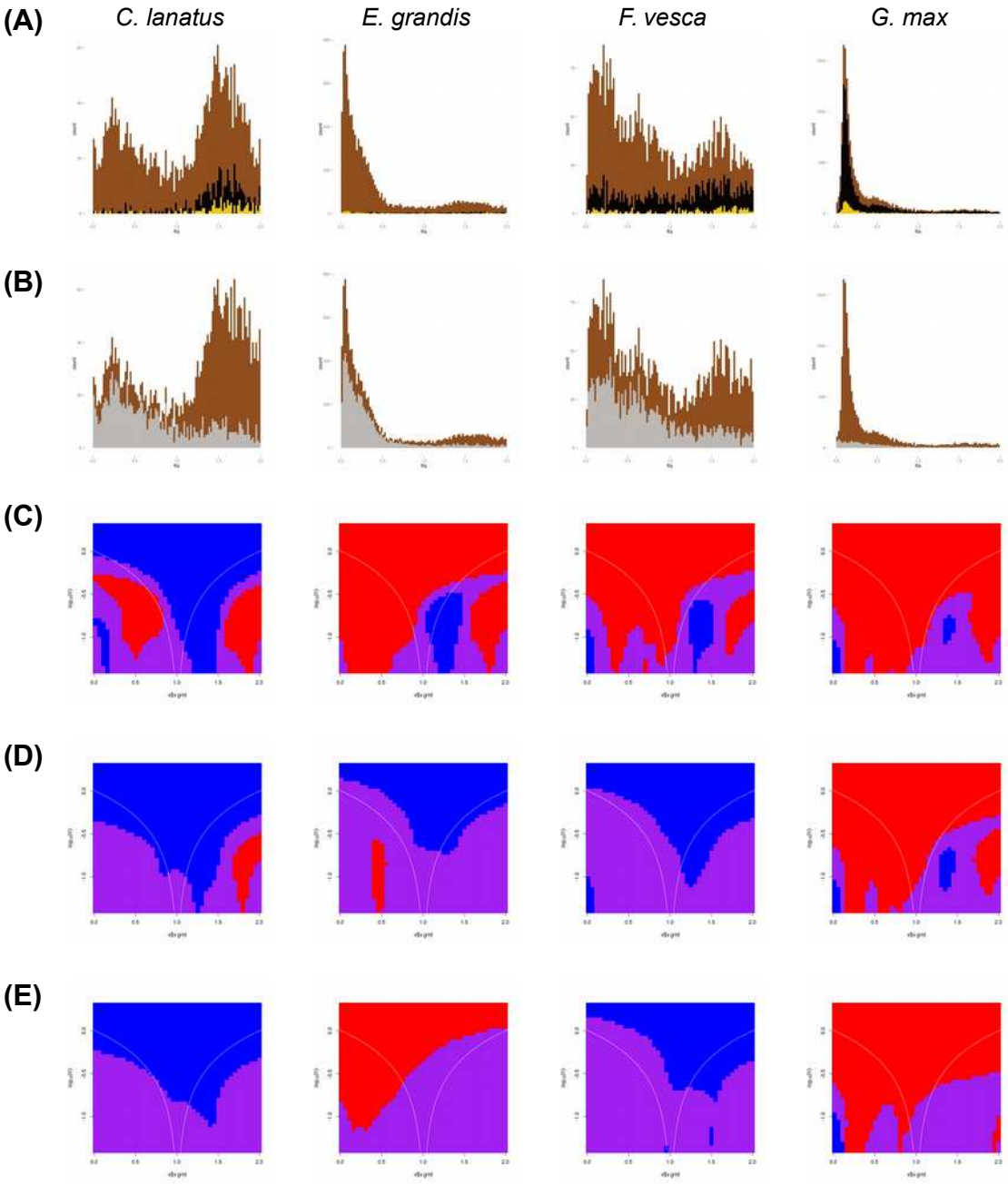
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**: 32–43.

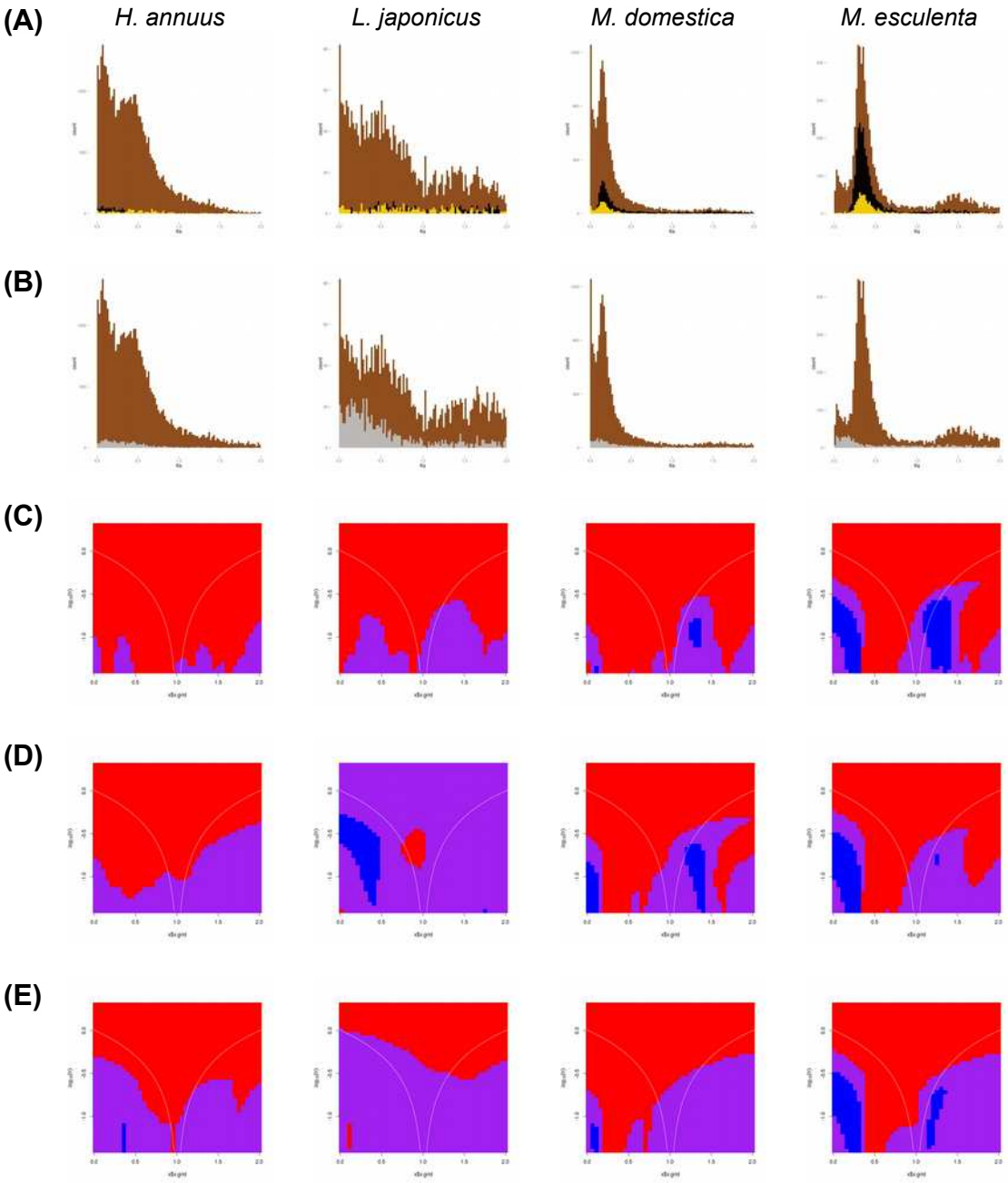
Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292–298.

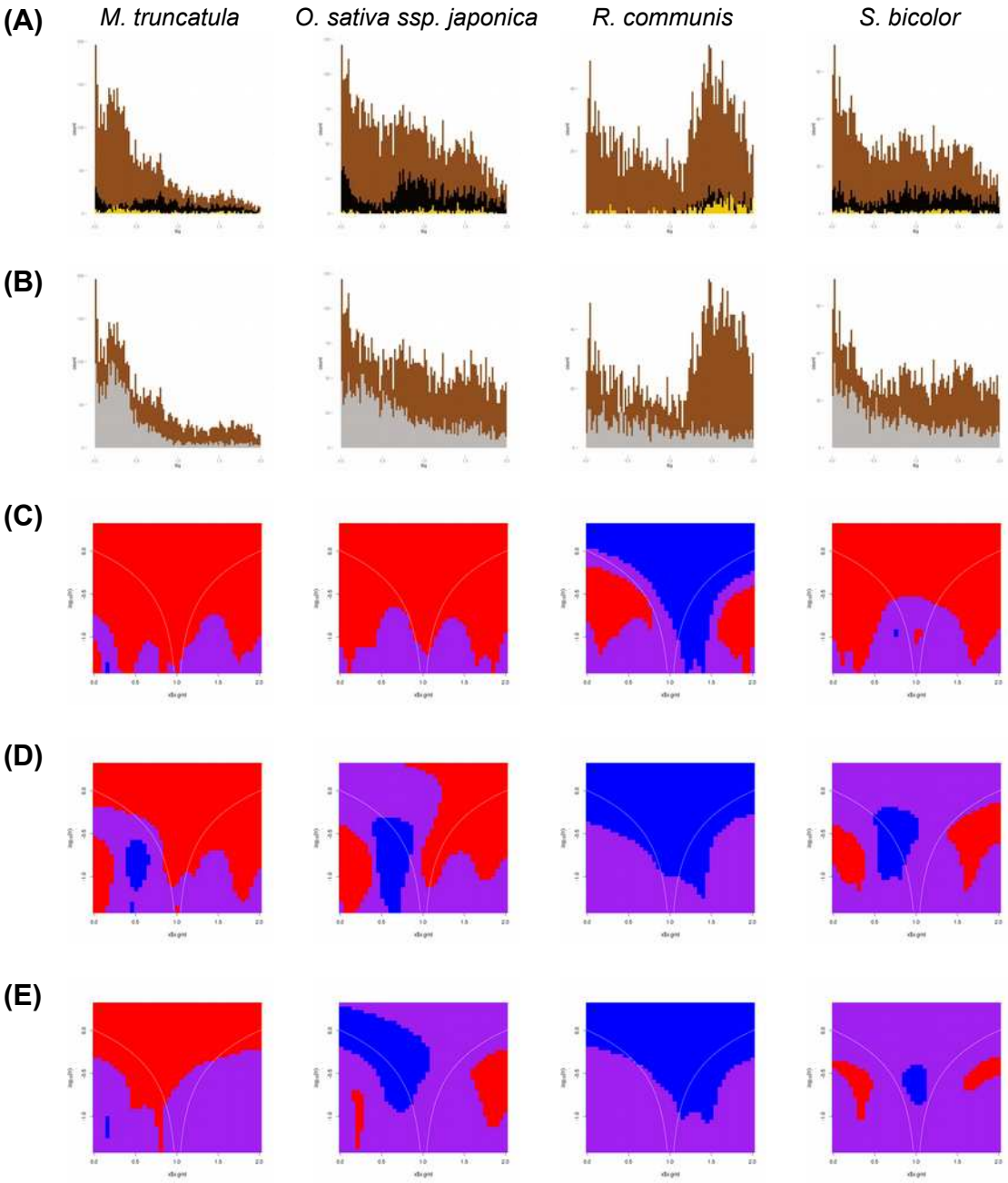
Both mechanism and age of duplications contribute to biased gene retention patterns in plants

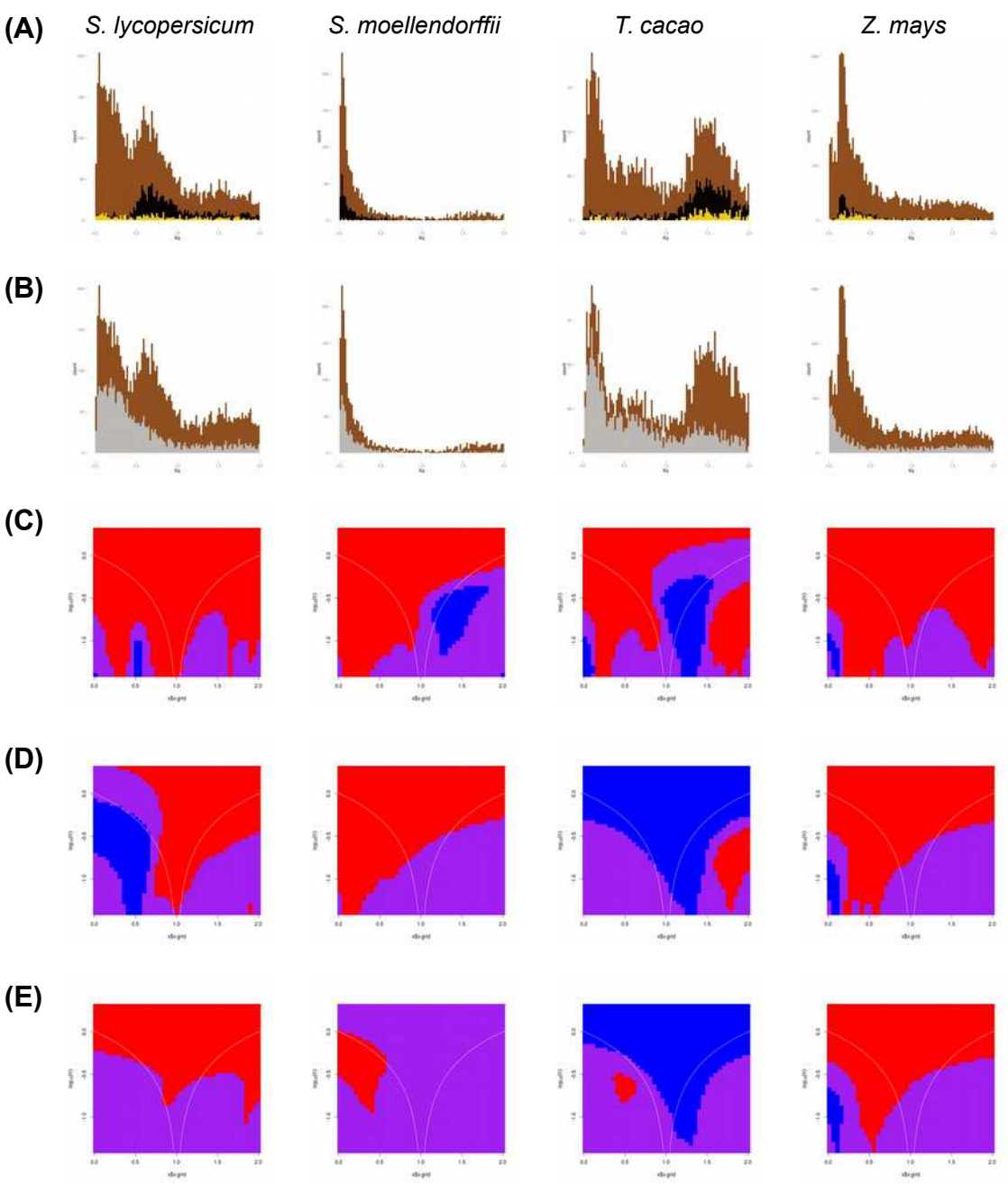
SUPPLEMENTAL MATERIAL


(A)
A. lyrata
A. trichopoda
B. distachyon
C. papaya
(B)
(C)
(D)
(E)


(A)
C. lanatus
E. grandis
F. vesca
G. max
(B)
(C)
(D)
(E)


(A)
H. annuus
L. japonicus
M. domestica
M. esculenta
(B)
(C)
(D)
(E)


(A)
M. truncatula
O. sativa ssp. japonica
R. communis
S. bicolor
(B)
(C)
(D)
(E)



Supplementary Figure S1. *Ks* age distributions of 20 species and correspondent maps. In (A) brown bars represent all duplicates (background), black and yellow represent the WGD-derived duplicates predicted by DAGchainer and transcription factors activity (GO:0003700) duplicates, respectively. (B) Brown bars represent the background and gray the tandem-derived duplicates. (C) SiZer maps of background. (D) SiZer maps of WGD-derived duplicates. (E) SiZer maps of transcription factors duplicates.

Supplemental Table 2. Number of duplicates annotated as transcription factor (TF) activity (GO:0003700) for 25 plant species, displayed by duplication type (predicted by the DAGchainer software) and grouped by *Ks* equivalent age ranges

Specie	TF Duplicates	TF Duplicates by *Ks* ranges											
		0 < Ks ≤ 0.5			0.5 < Ks ≤ 1			1 < Ks ≤ 1.5			1.5 < Ks ≤ 2		
		WGD	Tandem	Undefined	WGD	Tandem	Undefined	WGD	Tandem	Undefined	WGD	Tandem	Undefined
Arabidopsis lyrata	519	23	79	35	116	20	32	73	7	31	60	10	33
Arabidopsis thaliana	532	4	39	49	156	14	28	115	0	14	45	2	66
Amborella trichopoda	61	0	12	12	0	0	3	0	4	3	0	1	26
Brachypodium distachyon	127	3	8	5	9	12	8	9	14	15	13	17	14
Carica papaya	172	0	24	61	0	0	3	1	1	14	1	3	64
Citrullus lanatus	135	0	4	5	1	3	1	5	2	24	16	5	69
Eucalyptus grandis	261	0	95	35	2	17	9	4	8	26	3	8	54
Fragaria vesca	152	5	11	16	5	5	7	11	4	11	25	14	38
Glycine max	1224	579	52	291	95	28	23	49	15	13	53	19	7
Helianthus annuus	1124	2	23	556	2	7	370	0	3	76	0	3	82
Lotus japonicus	190	3	10	43	7	9	42	1	2	40	4	7	22
Malus domestica	1009	153	35	713	7	7	39	6	4	14	17	4	10
Manihot esculenta	668	211	8	275	38	4	44	4	4	21	15	2	42
Medicago truncatula	258	18	63	35	19	9	25	20	2	13	19	3	32
Oryza sativa ssp. indica	242	11	34	26	16	7	19	21	10	33	27	5	33
Oryza sativa ssp. japonica	187	5	15	5	18	6	16	19	7	33	29	4	30
Physcomitrella patens	102	1	2	10	5	2	42	3	1	26	1	0	9
Populus trichocarpa	692	434	26	129	12	4	9	15	6	10	29	7	11
Ricinus communis	139	0	5	7	0	2	7	4	0	25	13	4	72
Sorghum bicolor	145	9	20	8	7	5	10	14	8	21	15	9	19
Solanum lycopersicum	400	5	65	42	33	25	62	20	14	38	13	11	72
Selaginella moellendorffii	29	3	3	6	1	1	2	0	0	4	1	0	8
Theobroma cacao	209	5	21	4	0	8	5	21	9	35	47	7	47
Vitis vinifera	377	1	50	25	9	22	34	27	5	81	24	9	90
Zea mays	331	9	21	164	4	3	47	0	5	37	1	11	29

Supplemental Table 3. Gene Ontology (GO) categories overrepresented in duplicates with 1 < Ks ≤ 2 in nine or more of the 25 plant species analyzed in this study

Frequency	GO category	GO description	Species (abrev)
21	GO:0016020	membrane	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,lja,mes,mtr,osa,ptr,rco,sly,smo,tca,vvi,zma
19	GO:0071944	cell periphery	aly,ath,atr,cla,cpa,egr,fve,gma,ind,lja,mdo,mes,mtr,ppa,ptr,rco,sly,tca,vvi
18	GO:0022857	transmembrane transporter activity	aly,ath,atr,bdi,cpa,egr,fve,gma,ind,lja,mes,mtr,osa,ptr,sly,smo,tca,zma
18	GO:0015075	ion transmembrane transporter activity	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,mes,mtr,ptr,rco,sly,tca,vvi,zma
18	GO:0005886	plasma membrane	aly,ath,atr,cla,cpa,egr,fve,gma,ind,lja,mdo,mtr,osa,ptr,rco,sly,tca,vvi
18	GO:0005215	transporter activity	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,mes,mtr,osa,ptr,sly,smo,tca,zma
17	GO:0055085	transmembrane transport	aly,ath,atr,cla,egr,fve,gma,ind,lja,mdo,mes,mtr,ptr,sly,smo,tca,zma
17	GO:0051234	establishment of localization	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,lja,mtr,osa,ptr,sly,smo,tca
17	GO:0051179	localization	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,lja,mtr,osa,ptr,sly,smo,tca
17	GO:0006810	transport	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,lja,mtr,osa,ptr,sly,smo,tca
16	GO:1902578	single-organism localization	aly,ath,atr,bdi,cla,egr,fve,gma,ind,mes,mtr,osa,ptr,sly,smo,tca
16	GO:0044765	single-organism transport	aly,ath,atr,bdi,cla,egr,fve,gma,ind,mes,mtr,osa,ptr,sly,smo,tca
16	GO:0022892	substrate-specific transporter activity	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,ind,mes,mtr,ptr,sly,smo,tca
15	GO:0044699	single-organism process	aly,ath,atr,cpa,egr,fve,gma,ind,lja,mdo,mes,mtr,ptr,sly,tca
15	GO:0009987	cellular process	aly,ath,atr,cla,cpa,egr,fve,gma,ind,ptr,sbi,sly,smo,tca,vvi
15	GO:0009628	response to abiotic stimulus	aly,ath,atr,cla,cpa,egr,fve,gma,ind,mes,mtr,rco,sly,tca,vvi

14	GO:1901700	response to oxygen-containing compound	aly,atr,cla,cpa,egr,fve,gma,ind,mtr,ptr,rco,sly,tca,vvi
14	GO:0050896	response to stimulus	aly,ath,atr,cpa,egr,fve,gma,lja,mdo,ptr,rco,sly,tca,vvi
14	GO:0044763	single-organism cellular process	aly,ath,atr,cla,cpa,egr,fve,gma,ind,lja,ptr,sly,smo,tca
14	GO:0044464	cell part	aly,ath,atr,cla,cpa,egr,fve,ind,mtr,rco,sly,smo,tca,vvi
14	GO:0042221	response to chemical	aly,atr,cla,egr,fve,gma,ind,mdo,mtr,ptr,rco,sly,smo,tca
14	GO:0031224	intrinsic component of membrane	aly,ath,atr,bdi,cla,egr,fve,ind,mtr,osa,ptr,sly,smo,tca
14	GO:0022804	active transmembrane transporter activity	aly,ath,atr,egr,fve,gma,ind,lja,mtr,osa,ptr,sly,smo,zma
14	GO:0008509	anion transmembrane transporter activity	aly,ath,atr,egr,fve,gma,ind,mes,mtr,ptr,rco,sly,smo,tca
14	GO:0006811	ion transport	aly,ath,atr,bdi,cla,cpa,egr,fve,gma,mtr,ptr,rco,sly,tca
14	GO:0005623	cell	aly,ath,atr,cla,cpa,egr,fve,ind,mtr,rco,sly,smo,tca,vvi
14	GO:0005618	cell wall	aly,ath,atr,egr,fve,gma,ind,mdo,ppa,ptr,rco,sly,tca,vvi
13	GO:0065007	biological regulation	aly,ath,atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
13	GO:0050789	regulation of biological process	aly,ath,atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
13	GO:0022891	substrate-specific transmembrane transporter activity	aly,ath,atr,bdi,cla,egr,fve,gma,ind,mtr,ptr,sly,zma
13	GO:0006820	anion transport	aly,ath,atr,cla,egr,fve,gma,mes,mtr,ptr,rco,sly,tca
13	GO:0005975	carbohydrate metabolic	aly,ath,atr,egr,fve,gma,ind,mdo,mtr,osa,ptr,sbi,sly

		process	
12	GO:0040007	growth	aly,ath,atr,cla,egr,fve,gma,ind,rco,sly,tca,vvi
12	GO:0033993	response to lipid	aly,ath,atr,cla,cpa,egr,fve,gma,ind,sly,tca,vvi
12	GO:0031326	regulation of cellular biosynthetic process	atr,cla,egr,fve,hel,ind,mtr,osa,rco,sly,tca,vvi
12	GO:0030312	external encapsulating structure	aly,ath,atr,egr,fve,gma,ind,mdo,ppa,ptr,sly,vvi
12	GO:0010033	response to organic substance	atr,cla,egr,fve,gma,ind,mdo,mtr,rco,sly,tca,vvi
12	GO:0009719	response to endogenous stimulus	atr,cla,cpa,egr,fve,gma,ind,mtr,rco,sly,tca,vvi
12	GO:0003700	sequence-specific DNA binding transcription factor activity	atr,cla,egr,fve,hel,ind,mtr,osa,rco,sly,tca,vvi
12	GO:0001071	nucleic acid binding transcription factor activity	atr,cla,egr,fve,hel,ind,mtr,osa,rco,sly,tca,vvi
11	GO:2001141	regulation of RNA biosynthetic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:2000112	regulation of cellular macromolecule biosynthetic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:1903506	regulation of nucleic acid-templated transcription	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0097305	response to alcohol	aly,ath,atr,cla,cpa,fve,gma,ind,sly,tca,vvi
11	GO:0080090	regulation of primary	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi

		metabolic process	
11	GO:0060255	regulation of macromolecule metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0051252	regulation of RNA metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0051171	regulation of nitrogen compound metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0050794	regulation of cellular process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0048856	anatomical structure development	aly,atr,cla,egr,fve,ind,mtr,rco,sly,tca,vvi
11	GO:0044723	single-organism carbohydrate metabolic process	aly,ath,atr,cpa,egr,fve,gma,ind,mtr,osa,ptr
11	GO:0043229	intracellular organelle	atr,cla,egr,fve,hel,ind,rco,sly,smo,tca,vvi
11	GO:0043226	organelle	atr,cla,egr,fve,hel,ind,rco,sly,smo,tca,vvi
11	GO:0031323	regulation of cellular metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0019222	regulation of metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0019219	regulation of nucleobase-containing compound metabolic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0016021	integral component of	aly,ath,atr,egr,fve,ind,mtr,osa,ptr,sly,smo

		membrane	
11	GO:0015698	inorganic anion transport	aly,atr,cla,fve,gma,mtr,ptr,rco,sly,tca,zma
11	GO:0015291	secondary active transmembrane transporter activity	aly,ath,atr,fve,gma,mtr,ptr,sly,smo,tca,zma
11	GO:0015103	inorganic anion transmembrane transporter activity	aly,ath,atr,cpa,fve,gma,ind,mtr,ptr,rco,sly
11	GO:0010556	regulation of macromolecule biosynthetic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0010468	regulation of gene expression	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0009889	regulation of biosynthetic process	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0009791	post-embryonic development	atr,cla,cpa,egr,fve,ind,mtr,rco,sly,tca,vvi
11	GO:0009725	response to hormone	atr,cla,cpa,egr,fve,gma,ind,mdo,sly,tca,vvi
11	GO:0009505	plant-type cell wall	aly,ath,atr,egr,fve,gma,mdo,ptr,sly,tca,vvi
11	GO:0007154	cell communication	aly,atr,cpa,fve,gma,ind,lja,ptr,rco,sly,tca
11	GO:0006355	regulation of transcription, DNA-templated	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
11	GO:0005634	nucleus	atr,cla,egr,fve,hel,ind,osa,rco,sly,tca,vvi
10	GO:0071840	cellular component organization or biogenesis	aly,atr,cla,egr,ind,ppa,rco,sly,tca,vvi

10	GO:0071705	nitrogen compound transport	aly,ath,egr,fve,gma,mes,ptr,sly,tca,vvi
10	GO:0071702	organic substance transport	aly,ath,atr,egr,fve,gma,ind,osa,sly,tca
10	GO:0071555	cell wall organization	aly,ath,atr,cla,egr,gma,ptr,smo,tca,vvi
10	GO:0071495	cellular response to endogenous stimulus	aly,atr,cla,egr,fve,ind,mtr,sly,tca,vvi
10	GO:0071310	cellular response to organic substance	aly,atr,cla,egr,fve,ind,mtr,sly,tca,vvi
10	GO:0070887	cellular response to chemical stimulus	aly,atr,cla,egr,fve,ind,mtr,sly,tca,vvi
10	GO:0061458	reproductive system development	atr,cla,cpa,egr,fve,ind,mtr,rco,sly,tca
10	GO:0051716	cellular response to stimulus	aly,atr,cpa,fve,ind,lja,rco,sly,tca,vvi
10	GO:0048731	system development	atr,cla,egr,fve,ind,mtr,rco,sly,tca,vvi
10	GO:0048608	reproductive structure development	atr,cla,cpa,egr,fve,ind,mtr,rco,sly,tca
10	GO:0045229	external encapsulating structure organization	aly,ath,atr,cla,egr,gma,ptr,smo,tca,vvi
10	GO:0044767	single-organism developmental process	aly,atr,cla,egr,fve,ind,rco,sly,tca,vvi
10	GO:0044707	single-multicellular organism process	aly,atr,cla,egr,fve,ind,rco,sly,tca,vvi
10	GO:0044459	plasma membrane part	aly,ath,atr,egr,fve,ind,lja,rco,sly,tca
10	GO:0044425	membrane part	aly,ath,atr,cla,egr,fve,ind,sly,smo,tca

10	GO:0044424	intracellular part	atr,cla,egr,fve,ind,rco,sly,smo,tca,vvi
10	GO:0043231	intracellular membrane-bounded organelle	atr,cla,egr,fve,hel,ind,rco,sly,tca,vvi
10	GO:0043227	membrane-bounded organelle	atr,cla,egr,fve,hel,ind,rco,sly,tca,vvi
10	GO:0032870	cellular response to hormone stimulus	aly,atr,cla,egr,fve,ind,mtr,sly,tca,vvi
10	GO:0032502	developmental process	aly,atr,cla,egr,fve,ind,rco,sly,tca,vvi
10	GO:0032501	multicellular organismal process	aly,atr,cla,egr,fve,ind,rco,sly,tca,vvi
10	GO:0016772	transferase activity, transferring phosphorus-containing groups	atr,cpa,fve,hel,ind,lja,osa,ptr,sbi,zma
10	GO:0016043	cellular component organization	aly,atr,cla,egr,ind,ppa,ptr,smo,tca,vvi
10	GO:0009653	anatomical structure morphogenesis	aly,atr,cla,egr,fve,gma,ind,rco,sly,tca
10	GO:0009651	response to salt stress	aly,ath,atr,cpa,egr,fve,gma,ptr,rco,vvi
10	GO:0008324	cation transmembrane transporter activity	aly,ath,atr,cla,cpa,fve,gma,ptr,sly,tca
10	GO:0007275	multicellular organismal development	aly,atr,cla,egr,fve,ind,rco,sly,tca,vvi
10	GO:0006970	response to osmotic stress	aly,ath,atr,cpa,egr,fve,gma,ptr,tca,vvi
10	GO:0005622	intracellular	atr,cla,egr,fve,ind,rco,sly,smo,tca,vvi

10	GO:0001101	response to acid chemical	ath,atr,cla,cpa,egr,fve,gma,sly,tca,vvi
9	GO:0071554	cell wall organization or biogenesis	aly,ath,atr,cla,egr,gma,ppa,ptr,vvi
9	GO:0071396	cellular response to lipid	aly,atr,cla,egr,fve,gma,ind,sly,tca
9	GO:0048878	chemical homeostasis	aly,atr,cla,cpa,egr,fve,gma,ind,sly
9	GO:0048589	developmental growth	aly,ath,atr,cla,egr,fve,gma,ind,tca
9	GO:0048367	shoot system development	aly,ath,atr,cla,egr,fve,sly,tca,vvi
9	GO:0046873	metal ion transmembrane transporter activity	ath,atr,bdi,cla,cpa,fve,gma,ptr,sly
9	GO:0044700	single organism signaling	atr,cpa,fve,gma,ind,rco,sly,tca,vvi
9	GO:0043565	sequence-specific DNA binding	atr,cla,egr,fve,hel,ind,osa,tca,vvi
9	GO:0016773	phosphotransferase activity, alcohol group as acceptor	atr,cpa,fve,hel,lja,osa,ptr,sbi,zma
9	GO:0009605	response to external stimulus	atr,egr,fve,gma,ind,mdo,mtr,sly,vvi
9	GO:0007264	small GTPase mediated signal transduction	atr,cpa,fve,gma,lja,mdo,osa,rco,vvi
9	GO:0007165	signal transduction	atr,cpa,fve,gma,ind,rco,sly,tca,vvi
9	GO:0006796	phosphate-containing compound metabolic process	atr,cpa,fve,ind,lja,osa,ptr,sbi,zma
9	GO:0006793	phosphorus metabolic process	atr,cpa,fve,ind,lja,osa,ptr,sbi,zma

9	GO:0003824	catalytic activity	bdi,gma,ind,lja,mes,osa,ptr,sbi,zma
9	GO:0003006	developmental process involved in reproduction	aly,atr,cla,egr,fve,mtr,rco,sly,tca

Supplemental Table 4. Data source and abbreviation of the 25 plant genomes used in this study

Specie	Abbreviation	Release	Source
Arabidopsis lyrata	Aly	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Arabidopsis thaliana	Ath	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Amborella trichopoda	Atr	Plaza 3.0	ftp.psb.ugent.be/pub/plaza/plaza_public_dicots_03/
Brachypodium distachyon	Bdi	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Carica papaya	Cpa	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Citrullus lanatus	Cla	Plaza 3.0	ftp.psb.ugent.be/pub/plaza/plaza_public_dicots_03/
Eucalyptus grandis	Egr	Plaza 3.0	ftp.psb.ugent.be/pub/plaza/plaza_public_dicots_03/
Fragaria vesca	Fve	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Glycine max	Gma	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Helianthus annuus	Hel	HA412 v1.1	http://www.sunflowergenome.org/
Lotus japonicus	Lja	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Malus domestica	Mdo	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Manihot esculenta	Mes	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Medicago truncatula	Mtr	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Oryza sativa ssp. indica	Ind	Plaza 3.0	ftp.psb.ugent.be/pub/plaza/plaza_public_monocots_03/
Oryza sativa ssp. japonica	Osa	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Physcomitrella patens	Ppa	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Populus trichocarpa	Ptr	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Ricinus communis	Rco	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Sorghum bicolor	Sbi	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Solanum lycopersicum	Sly	Plaza 3.0	ftp.psb.ugent.be/pub/plaza/plaza_public_dicots_03/
Selaginella moellendorffii	Smo	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Theobroma cacao	Tca	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Vitis vinifera	Vvi	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/
Zea mays	Zma	Plaza 2.5	ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/

CONCLUSÕES GERAIS

CONCLUSÕES GERAIS

O destino evolutivo e a retenção de genes duplicados nos genomas parecem ser moldados por peculiaridades inerentes a cada organismo poliplóide e sua constante busca por adaptação, onde as características como o mecanismo de duplicação, função do gene, fatores epigenéticos e ambientais podem temporariamente exercer maior ou menor influência.

Parálogos das duas famílias de genes alvos do nosso estudo, metionina sintase a oligossacarídeos de rafinose, evoluíram sob diferentes restrições em soja. As funções originais de cada gene são preservadas em ao menos um dos parálogos, onde mutações ocorrem apenas em sítios específicos da proteína e eventualmente são fixadas através de seleção positiva.

Duas difereças evolutivas marcantes foram encontradas entre Monocotiledôneas e Dicotiledôneas para as duas famílias de genes supracitadas. Monocotiledôneas carecem da isoforma cloroplastídica da metionina sintase, e sua estaquiose sintase não possui uma inserção de aproximadamente 80 aminoácidos presentes em todas as espécies Dicotiledôneas e também *Amborella trichopoda*. Apesar do distanciamento evolutivo entre as espécies, a função original dos genes parecem ter sido apenas especializadas ou adaptadas. Interessante notar que o gene da estaquiose sintase de soja é de cópia única e evoluiu sob forte seleção purificadora.

Algumas categorias de genes, como fatores de transcrição, possuem suas cópias de genes de origem mais antigas superrepresentadas em seus genomas em detrimento das cópias que foram geradas em eventos de duplicação mais recentes, independentemente do mecanismo de duplicação pelo qual as cópias foram originadas. Essa observação pode apenas ser conciliada com a Hipótese do Equilíbrio de Genes se as restrições de dosagem houverem sido mais intensas durante as duplicações de genoma inteiro mais antigas. O decorrer de seguidos eventos de duplicação e o consequente aumento do número de parálogos podem ter levado ao relaxamento da seleção purificadora, consequentemente culminando no padrão de retenção tendenciosa dos parálogos de origem antiga observados neste trabalho.