RECOMBINATION OF chl-*fus* GENE (PLASTID ORIGIN) DOWNSTREAM OF *hop*: A LOCUS OF CHROMOSOMAL INSTABILITY

LIBIA CATALINA SALINAS CASTELLANOS

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGÍA
BUCARAMANGA
2015

RECOMBINATION OF chl-*fus* GENE (PLASTID ORIGIN) DOWNSTREAM OF *hop*: A LOCUS OF CHROMOSOMAL INSTABILITY

LIBIA CATALINA SALINAS CASTELLANOS

Trabajo de grado presentado como requisito para optar al título de Bióloga

Director
DR. JORGE HERNÁNDEZ-TORRES

Co-Director
DR. JACQUES CHOMILIER

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGÍA
BUCARAMANGA
2015

AGRADECIMIENTOS

A mis Padres Libia y Alfredo, por el amor, la comprensión e inagotable apoyo. A Caro, mi hermana y cómplice por la motivación y compañía, eres mi motor. A Silvi mi hermana adoptiva, por la dedicación y paciencia. Los amo.

A mis profesores Jorge Hernández Torres y Jacques Chomilier, por el apoyo, la confianza en mi trabajo y la capacidad para guiar mis ideas; han sido un aporte invaluable, no solamente en el desarrollo de ésta tesis, sino también en mi formación como investigadora.

A mis Amigos, por los buenos momentos, por la compañía durante todos éstos años y enseñanzas en mi formación académica y personal.

TABLE OF CONTENTS

Pag

LIST OF FIGURES

Pag.

LIST OF TABLES

Pag.

ADDITIONAL FILES

Pag

RESUMEN

TÍTULO: RECOMBINATION OF chl-*fus* GENE (PLASTID ORIGIN) DOWNSTREAM OF *hop*: A LOCUS OF CHROMOSOMAL INSTABILITY*.

AUTOR: LIBIA CATALINA SALINAS CASTELLANOS†

PALABRAS CLAVES: Proteínas TPR; gen *hop*; cEF-G; gen chl-*fus*; Microsintenia; Exon shuffling; Fase de los intrones

DESCRIPCION:

La co-chaperona Hop actúa como adaptadora en el plegamiento y maduración de Hsp70 y Hsp90. El gen *hop* es de origen eucariótico. Por otra parte, el factor de elongación cloraplástico G (cEF-G) cataliza el paso de translocación en la síntesis de proteínas en el cloroplasto. El gen chl-*fus* es de origen cloroplástico. Ambas proteínas fueron originadas por la duplicación de dominios. Fue demostrado, que el gen nuclear chl-*fus*, el cual codifica para la proteína cEF-G, se localiza en transcipción convergente con el gen *hop* en *Glycine max*. Analizamos 51 genomas vegetales desde Chlorophyta hasta plantas superiores, para determinar sí el gen chl-*fus* fue transferido directamente a la célula Proto-eucariota con *hop.* Ambos genes vienen de eventos de exón/modulo duplicación, exploramos la participación de los intrones en el origen y los consecutivos cambios en la estructura de los genes. Reconstruimos la historia evolutiva de los dos genes de transcripción convergente en plantas, basado en su estructura, microsintenia y microcolinealidad. A pesar del alto grado de microcolinealidad (65%) se demostró que su continuidad es producto de arreglos cromosómicos. Según, la predicción de la estructura exón-intrón se infirieron los eventos moleculares que dieron lugar a los genes actuales. Se propuso un modelo por recombinación donde los intrones fase-0 fueron esenciales para la duplicación de los dominios y los intrones fase-1 para el reclutamiento del péptido de tránsito. Finalmente, demostramos la susceptibilidad natural de la región intergénica a recombinarse o perderse, afectando seriamente la integridad del gen chl-*fus* en el futuro. Concluimos que el gen chl-*fus* fue transferido desde el cloroplasto a un cromosoma diferente al de *hop* en la célula eucatiótica fotosintética primitiva. Antes de aparecer en plantas superiores fue recombinado junto a *hop.* La recombinación mediada por la simetría de los intrones fue esencial para la evolución de los genes.

* Proyecto de investigación

† Facultad de Ciencias Básicas. Escuela de Biología. Director Dr. Jorge Hernández Torres. Co-Director Dr. Jacques Chomilier

ABSTRACT

TITLE: RECOMBINATION OF chl-*fus* GENE (PLASTID ORIGIN) DOWNSTREAM OF *hop*: A LOCUS OF CHROMOSOMAL INSTABILITY[‡]

AUTHOR: LIBIA CATALINA SALINAS CASTELLANOS[§]

KEYWORDS: TPR proteins; *hop* gene; cEF-G; chl-*fus* gene; Microsynteny; Exon shuffling; Intron phase

DESCRIPITON:

The co-chaperone Hop has been shown to act as an adaptor for protein folding and maturation, in concert with Hsp70 and Hsp90. The *hop* gene is of eukaryotic origin. Likewise, the chloroplast elongation factor G (cEF-G) catalyzes the translocation step in chloroplast protein synthesis. The chl-*fus* gene is of plastid origin. It was demonstrated that the nuclear chl-*fus* gene, which encodes the cEF-G protein, locates in opposite orientation to a *hop* gene in *Glycine max*. We explored fifty-one available plant genomes from Chlorophyta to higher plants, to determine whether the chl-*fus* gene was transferred directly downstream of the primordial *hop* in the proto-eukaryote host cell. Both genes came from exon/module duplication events. We reconstructed the evolutionary history of the two convergent plant genes, on the basis of their gene structure, microsynteny and microcolinearity. Despite a high degree (65%) of microcolinearity among vascular plants, our results demonstrate that their adjacency was a product of chromosomal rearrangements. Based on predicted exon−intron structures, we inferred the molecular events giving rise to the current form of genes. Therefore, we propose a simple model of exon/module shuffling by intronic recombinations in which phase-0 introns were essential for domain duplication, and a phase-1 intron for transit peptide recruiting. Finally, we demonstrate a natural susceptibility of the intergenic region to recombine or delete, seriously threatening the integrity of the chl-*fus* gene for the future. Our results are consistent with the interpretation that the chl-*fus* gene was transferred from the chloroplast to a chromosome different from that of *hop* in the primitive photosynthetic eukaryote. Before the appearance of higher plants, it was recombined downstream of *hop*. Exon/module shuffling mediated by symmetric intron phases was essential for gene evolution. The intergenic region is prone to recombine, risking the integrity of both genes.

[‡] Research Proyect

[§] Science Faculty. School of Biology. Director Dr. Jorge Hernández Torres. Co-Director Dr. Jacques Chomilier

INTRODUTION

Conserved synteny is the degree to which genes remain on corresponding chromosomes [1,2]. The analysis of conserved microsynteny (i.e., small regions of synteny) is a useful method to unveil the molecular events that have occurred since the transfer of organellar genes to the nucleus. To unravel the details of genome recombination during speciation and that are associated with the formation of new species, conserved microsynteny analysis is also essential. Otherwise, gene colinearity is the conservation of gene content and orders over time [1]. The study of how gene orders are conserved reveals the degree of chromosome rearrangement within specific genomes. In this work, we describe the evolutionary history of two convergent plant transcription genes, *hop* and chl-*fus*. We examined the gene microsynteny and microcolinearity of the pair *hop* (nuclear origin) – chl-*fus* (chloroplast origin) from fifty-one plant nuclear genomes, describe their phylogenetic relationships, and discuss the influence of intron phase distribution on the evolution of both genes by exon shuffling. Predicted recombination events, in higher plants, support the hypothesis that the chromosomal region downstream of the *hop* gene is prone to recombine, having favored the shuffling of the chloroplast chl-*fus* gene adjacently to *hop*, in an opposite orientation.

The co-chaperone Hop [*h*eat shock protein (HSP) *o*rganizing *p*rotein] has been shown to bind both Hsp70 and Hsp90 into supercomplexes that act as an adaptor for protein folding and maturation [3]. The Hop protein is composed of three TPR domains: TPR1 is followed by one DP domain and then one Ch. AA domain; TPR2A; and TPR2B, which is followed by one DP domain [4,5]. Previous analyses of human and mouse genomes suggest that *hop* genes result from successive duplication of an ancestral TPR–DP module surrounded by introns of the same phase [6]. Hop is a ubiquitous eukaryotic protein, implying that its evolutionary origin dates back to the emergence of the first eukaryotic cells [7]. Furthermore,

molecular and bioinformatic studies conclude that Hop is encoded by orthologous gene families in all eukaryotes [6]. The *hop* gene is also found in plants; one member of the family was found downstream in convergent transcription with the chl-*fus* gene, which encodes the chloroplast-specific translation elongation factor G (cEF-G) [8].

According to the endosymbiotic theory, chloroplasts and mitochondria arose from the engulfment of prokaryotic cells by a proto-eukaryotic cell. Through evolutionary time, around 14-20% of genes of chloroplast genome origin were transferred to the nucleus [9-11]. As a consequence, the transferred genes had to adapt to the nuclear genetic system (i.e., eukaryotic promoters, spliceosomal introns, etc.). Nuclear-encoded chloroplast proteins that are synthesized in the cytosol are imported through the outer and inner envelope membranes of chloroplast; this is possible because transferred genes recruited DNA sequences coding for an N-terminal transit peptide [12]. From the sequencing of the first plastid genomes e.g., *Nicotiana tabacum* [13], *Marchantia polymorpha* [14], *Oryza sativa* [15], *Euglena gracilis* [16], it was concluded that the chl-*fus* gene is no longer located in the chloroplast but strictly found in the nucleus [17]. The first plant chl-*fus* gene was cloned and sequenced from *Glycine max*; the gene is split three times by introns of 330, 508 and 288 bp [18]. The first exon codes for a typical chloroplast transit peptide that must be removed after translocation into the stroma [12]. The EF-G protein is also a ubiquitous protein; it is of prokaryote origin and found in bacteria, chloroplasts, and mitochondria [19,20]. The EF-G catalyzes the translocation step in prokaryote-type protein synthesis and also promotes ribosome disassembly together with ribosome recycling factor (RRF) [21]. Surprisingly, near to nothing has been published about the plant chl-*fus* gene, since it was cloned and sequenced in *G. max* [18].

The microcolinearity between *hop* and chl-*fus* genes in *G. max* raises many interesting questions: Are all *hop* and chl-*fus* plant genes arranged in a convergent orientation, as in *G. max* (microcolinearity)? Was chl-*fus* directly transferred from

chloroplasts, downstream of the primordial *hop*? If that were the case, would be possible to explain, based on sequence analysis, why the chl-*fus* gene was transferred and functionally established downstream of *hop*? In vertebrates, the *hop* gene is organized in recombinable modules TPR−DP, surrounded by introns of the same phase. This could explain the evolutionary origin of *hop* by triplication of an ancient unit TPR−DP. Does the exon–intron organization of plant *hop* genes support this hypothesis? And finally, how can the study of the pair of genes *hop* and chl-*fus* contribute to the understanding of the evolution of plant genomes? Here, all these questions are discussed and, on the basis of the findings, models for the evolution of *hop* and chl-*fus* genes are proposed.

1. METHODS

1.1. ACCESSION NUMBERS AND EXON ASSEMBLY

The *Glycine max* chl-*fus* gene [GenBank: X71439] [18] was used as query sequence for BLAST searches in Genbank [22]. Accession numbers of retrieved contigs are in Table 1. Exon assembly was resolved using Geneious software [23] combined with manual adjustments. *G. max* cEF-G [18] and human Hop [24] were used as reference for exon assembly and protein domain definition. cDNAs from *A. thaliana* (cv. Columbia) were: cDNA1 [GenBank:BX815512], cDNA2 [GenBank:AK228637] and cDNA3 [GenBank:NM_104952] for *hop* gene and cDNA1[GenBank:NM_104951], cDNA2 [GenBank:AK221774] and cDNA3 [GenBank:AY142646] for chl-*fus* gene.

1.2. INTRON PHASE DEFINITION

Intron phase was assigned as stated by Patty [25]. Phase-0 introns split the open reading frame (ORF) within two codons, e.g., 5'GGC **CAG**:GT— intron— AG:**GTC** ACG3'. Phase-1 introns split the ORF between the first and second nucleotides of a codon, e.g., 5'GA GCA **G**:GT—intron—**AG**:GT CAC G3'. Phase-2 introns interrupt the ORF between the second and third nucleotides of a codon, e.g., 5'G AGC **AG**:GT—intron—AG:**G** TCA TG3'. Recombinable modules are defined as a set of exons flanked by introns of the same phase, typically phase-0 [6].

1.3. PHYLOGENETIC ANALYSIS

Maximum Likelihood phylogenetic trees were constructed using RaxML program version 7.3.0 [26]. All other settings were left as default, with 1000 replicates for

bootstrapping. Human Hop protein [GenBank:NP_006810] and *A. thaliana* mEF-G [GenBank:NC_003070] were used as outgroups. Additional EF-G sequences were: *A. caulinodans* ORS 571 [GenBank:YP_001525473], *A. fabrum* str. C58 [GenBank:NP_354925], *F. alni* ACN14a [GenBank:YP_711337], *K. radiotolerans* [GenBank: SRS30216YP_001360437], *R. prowazekii* str. Madrid E [GenBank: NP_220524], *Synechococcus* sp. [GenBank: P18667] and *S. coelicolor* [GenBank: NP_628821].

1.4. HYDROPHOBIC CLUSTER ANALYSIS (HCA)

Through the HCA method [27], we circumscribed the TPR and DP domains of orthologous Hop proteins. Besides, protein alignments were performed by this method. HCA is a method of protein analysis, implying the representation of amino acid sequences into a 2D space. The image is duplicated to exhibit the neighboring residues for each amino acid. Hydrophobic amino acids form clusters that correspond to the centers of regular secondary structures [28]. The shapes of the clusters are a keen indication of the nature of the secondary structure [29]. Clusters are roughly vertical when they code for a strand, while helixes are fairly horizontal. In a 2D protein alignment, the conserved shapes of the clusters are more important than the exact conservation of the residues inside the clusters. Thus, HCA allows alignments between very distantly related proteins, with as low as 10% identity. Additional sequences used in HCA alignments were: *O. lucimarinus* [GenBank:NC_009360], *Micromonas* sp. RCC299 [GenBank:NC_013040], *C. reinhardtii* [GenBank:NW_001843572], *L. alabamica* [GenBank:ASXC01000179], *A. arabicum* [GenBank:ASZG01007785] and human [GenBank:NC_000011].

2. RESULTS

2.1 CAPTURE AND VALIDATION OF PLANT *hop* AND chl-*fus* GENE SEQUENCES

The first chl-*fus* gene was cloned and characterized in *G. max* [18]. Protein sequence alignments of the encoded open reading frame (ORF), as well as the chloroplast-type transit peptide analysis, suggested that the mature protein belongs to the chloroplast protein synthesis machinery [18,30]. For example, the *Arabidopsis thaliana* cEF-G (At_cEF-G) shares 44% identity with its mitochondrial counterpart (At_mEF-G), while 59% with *Escherichia coli* EF-G (γ-Proteobacteria), 54% with *Synechococcus* sp. EF-G (Cyanobacteria) and 62% with *Agrobacterium fabrum* (α-Proteobacteria) EF-G.

Gene mapping efforts in *G. max,* following the discovery of chl-*fus* gene, revealed that chl-*fus* locates downstream of *hop* gene in an opposite orientation [8]. Thus, BLAST searches against other plant genomes, using the *G. max* chl-*fus* sequence as query, could be used to retrieve chl-*fus* orthologs and therefore *hop* genes from new plant sequenced genomes. Microsynteny analyses of such new genomes would help us to determine the ubiquitously of the transcriptional convergence of *hop* and chl-*fus* genes, or if *G. max* is an isolated case. We then used the *G. max* chl-*fus* gene as a BLAST query sequence to search for plant genomic contigs, coding for a predicted cEF-G preceded by a chloroplast-type transit peptide [18], concurrently with a *hop* gene in convergent transcription. The families, genera and species, and corresponding accession numbers of retrieved contigs obtained from Genbank are provided in Table 1. In plant species whose chl-*fus* and *hop* genes were not syntenic, the *G. max hop* gene was used as query to capture Hop encoding sequences. Using the *G. max* chl-*fus* and *hop* genes as reference, we mapped the predicted exon–intron structure of each gene for all plant species.

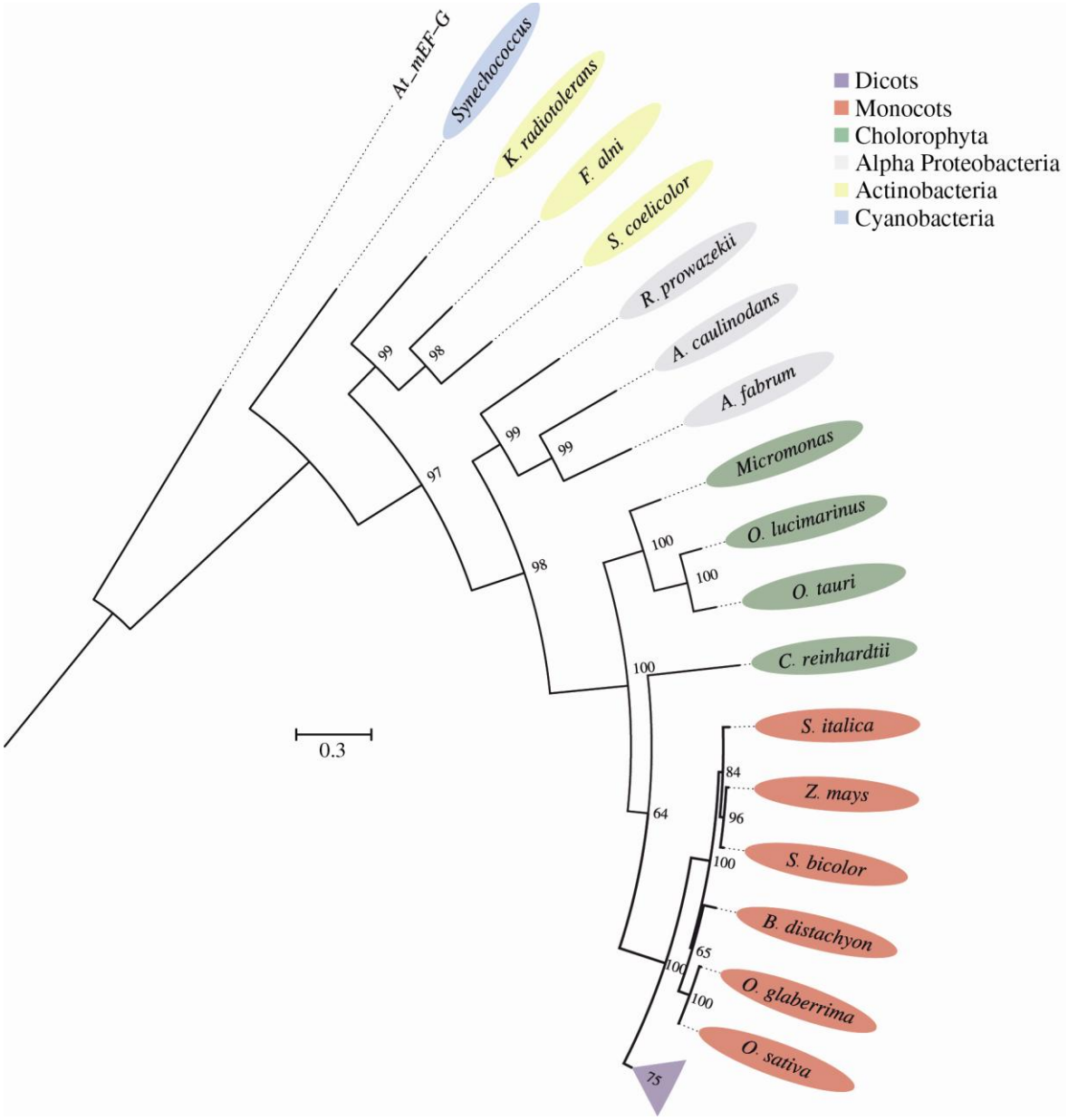
Table 1. Accession numbers of retrieved contigs sequences obtained from plant genome databases. The number of introns of *hop* and chl-*fus* genes, respectively, is given in arabic numbers.

Family	Species	Introns	Accession numbers
CHLOROPHYTA			
Mamiellaceae	*Micromonas sp. RCC299*	1-1	XP_002500383; XP_002500081
	Ostreococcus lucimarinus	0-1	XP_001418158; XP_001419031
	Ostreococcus tauri	0-1	XM_003079642; XM_003080500
Chlamydomonadaceae	*Chlamydomonas reinhardtii*	12-9	XP_001691869; XM_001701793
MONOCOTS			
Musaceae	*Ensete ventricosum*	6-3	AMZH01008475; AMZH01015354
Poaceae	*Brachypodium distachyon*	6-3	NC_016135
	Oryza glaberrima	6-3	ADWL01008993
	Oryza sativa	6-3	CM000129
	Setaria italica	6-3	NW_004675967
	Sorghum bicolor	6-3	NC_012875
	Zea mays	6-3	GK00032
Arecaceae	*Elaeis guineensis*	6-3	ASJS01002389-94
	Phoenix dactylifera	6-3	ATBV01012962
DICOTS			
Cucurbitaceae	*Citrullus lanatus*	6-3	AGCB01004585; AGCB01006484
	Cucumis melo	6-3	CAJI01012439; CAJI01003926
	Cucumis sativus	6-3	XM_004147890; XM_004147564
Cannabaceae	*Cannabis sativa*	6-3	AGQN01077260
Moraceae	*Morus notabilis*	6-3	ATGF01007958
Rosaceae	*Fragaria vesca subsp vesca*	6-3	NC_020495
	Malus domestica	6-3	ACYM01058960
	Prunus mume	6-3	AOHF01010810
	Prunus persica	6-3	AEKV01005456
	Pyrus x bretschneideri	6-3	AJSU01026097
Fabaceae	*Cajanus cajan*	6-3	AGCT01009484-85
	Cicer arietinum	6-3	XM_00451602; XM_004515686
	Glycine max	6-3	XP_003549898
	Lupinus angustifolius	6-3	AOCW01121688; AOCW01054016
	Medicago truncatula	6-3	NC_016411; NC_016410
Euphorbiaceae	*Hevea brasiliensis*	6-3	AJJZ010763885
	Jatropha curcas	6-3	BABX02001448
	Ricinus communis	6-3	NW_002994274
Linaceae	*Linum usitatissimum*	6-3	AFSQ01027627-29
Salicaceae	*Populus trichocarpa*	6-3	NC_008469
Malvaceae	*Gossypium raimondii*	6-3	AMOP01022205
	Theobroma cacao	6-3	CACC01007881
Brassicaceae	*Aethionema arabicum*	5-3	ASZG01007785
	Arabidopsis lyrata	6-3	NW_003302554
	Arabidopsis thaliana	6-3	NC_003070
	Brassica rapa	6-2	AENI01007476
	Capsella rubella	6-3	ANNY01000463
	Eutrema parvulum	6-3	AFAN01000006
	Eutrema salsugineum	6-3	AHIU01002482
	Leavenworthia alabamica	6-3	ASXC010000179
	Sisymbrium irio	6-3	ASZH01019437
Caricaceae	*Carica papaya*	6-3	ABIM01007984
Rutaceae	*Citrus sinensis*	6-3	AJPS01000059
Vitaceae	*Vitis vinifera*	6-3	AM459130
Solanaceae	*Nicotiana sylvestris*	6-3	ASAF01010839-40
	Nicotiana tomentosiformis	6-3	ASAG01110979
	Solanum lycopersicum	6-3	AP009300
	Solanum tuberosum	6-3	AEWC01024049

We show in Figure 1 a well-supported phylogenetic tree constructed with EF-G sequences from Actinobacteria, α-Proteobacteria and Cyanobacteria and 51 cEF-G sequences from Chlorophyta, Monocots and Dicots. The branching pattern of the cladogram indicates that EF-Gs from all life forms descended from a common ancestor. According to the evolutionary relationships, plant cEF-G sequences group together in a single branch with *G. max* cEF-G (our reference sequence), confirming that the assembled plant ORFs belong all to the chloroplast EF-G family. Chlorophyta cEF-G sequences share a common ancestor with higher plants, excepting *Chlamydomonas reinhardtii*, which appear to be closely related to higher plants than the other members of green algae. Monocot and dicot branches are coherent with canonical evolutionary trees; however, dicot branch had low support (bootstrap values less than 50%) resulting in this clade being unresolved [31]. As already reported [20], cEF-G sequences show more identity with α-proteobacterial EF-G than with cyanobacteria and this finding is confirmed in Figure 1, without exception. Taking these results together, we concluded that retrieved cEF-G sequences from Genbank were correctly reconstructed and they code for the chloroplast translation elongation factor G.

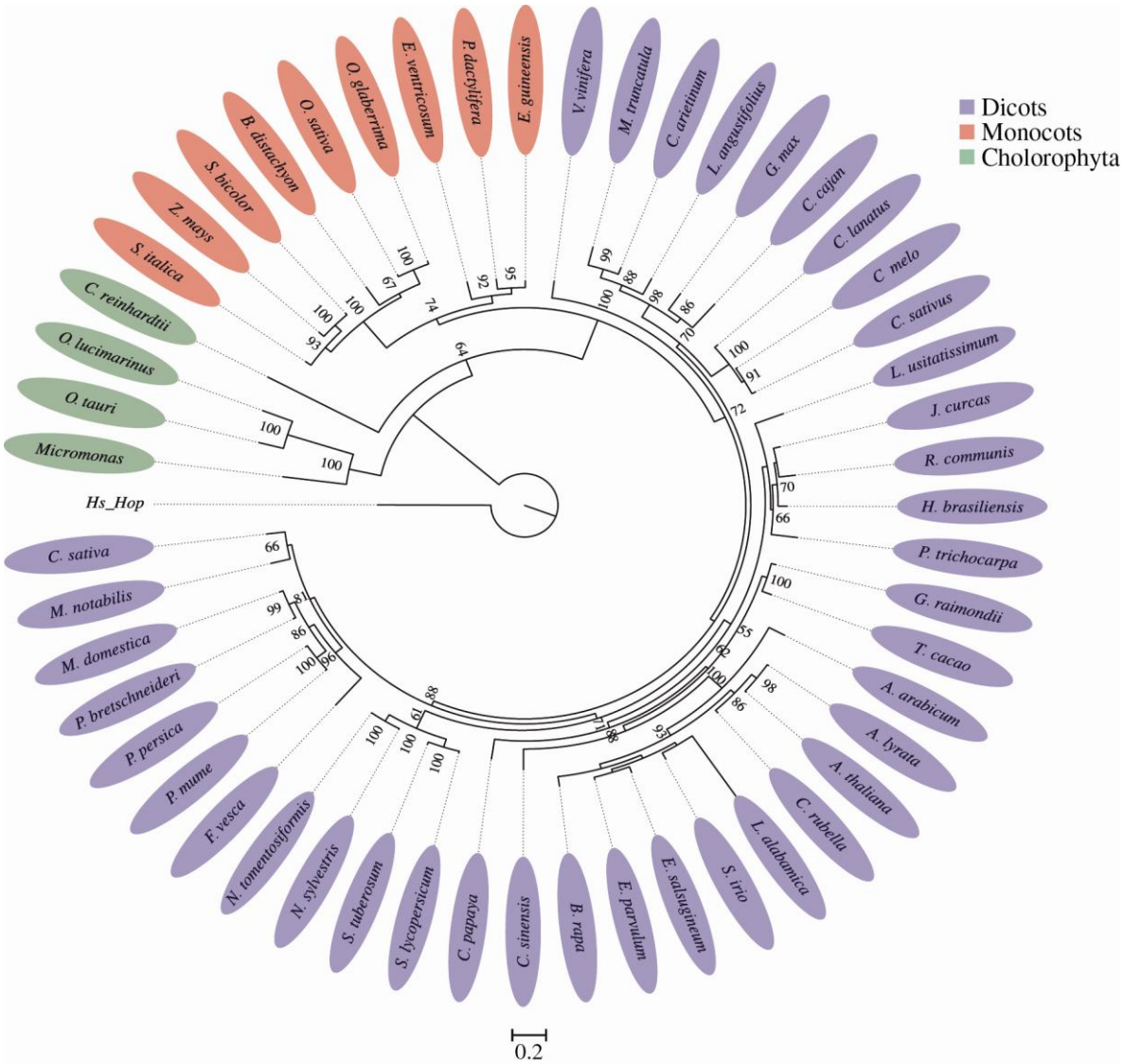
After intron removal from *hop* genes, the reconstructed Hop sequences were used to build a second phylogenetic tree (Figure 2). As expected, the assembled ORFs belong all to the plant Hop family which exhibit a large amount of divergence with respect to the outgroup (Human Hop). As seen in Figure 2 the inferred relationships among these protein sequences are robust and all branches are well supported, coherently with current plant systematics.

Figure 1. Phylogenetic tree of chloroplast elongation factor cEF-G sequences from 51 plant genomes. Bootstrap values are in Arabic numbers. Dicot branch (purple) was collapsed (bootstrap values less than 50%). Chlorophyta species are in green, while monocots are pink. Other members of the EF-G family: At_mEF-G: *A. thaliana* mitochondrial elongation factor G (outgroup). Cyanobacterial cEF-G (in Blue): *Synechococcus*. Actinobacterial EF-G (in yellow): *K. radiotolerans*, *F. alni* and *S. coelicolor*. α-Proteobacterial EF-G (in gray): *R. prowazekii*, *A. caulinodans* and *A. fabrum*. 0.3: Distance scale.



Interestingly, *Leavenworthia alabamica* grouped with the other members of Brassicales but with an unusual long evolutionary distance (Figure 2). Exceptionally, *L. alabamica* contains three tandem repetitions of the VPEVEKKLEPEPEP motif within the Ch. AA domain, while all other plants possess only one. These results confirm the correct assembly of *hop* genes from retrieved contigs.

Figure 2. Phylogenetic tree of Hop protein sequences of 51 plant genomes. Chlorophyta species are in green, monocots are in pink, and dicots are in purple. Hs_Hop: Human Hop protein (outgroup). 0.2: Distance scale.



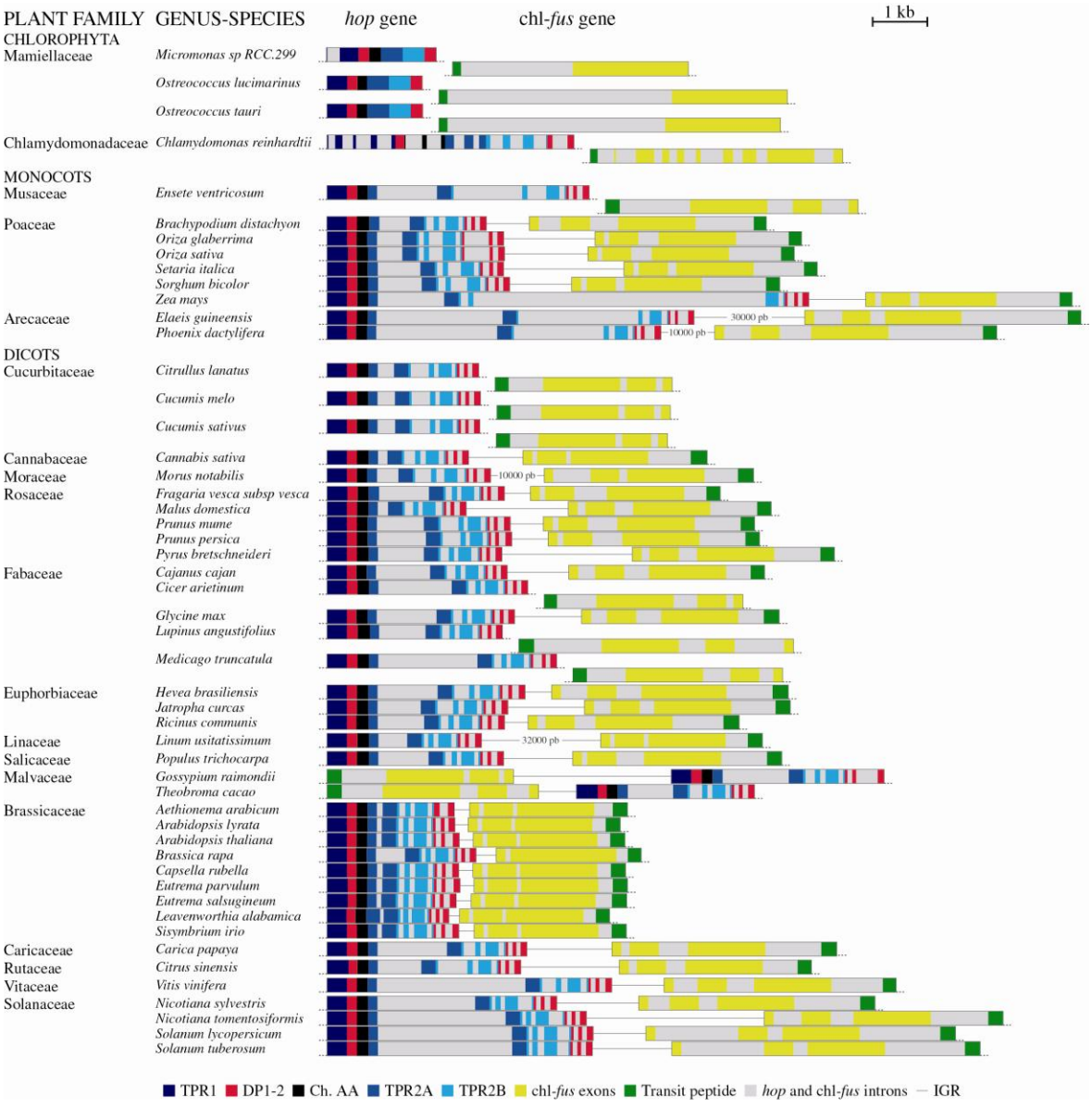
2.2 PRESERVED MICROSYNTENY AND MICROCOLINEARITY BETWEEN *hop* AND chl-*fus* GENES

The *hop* and chl-*fus* genes were discovered in *G. max* one after the other on the same chromosome, in convergent transcription arrangement [8]. This finding leads to two intriguing evolutionary questions: Have *hop* and chl-*fus* genes been together from the first to the present-day photosynthetic eukaryotes? Or, is their chromosomal contiguity strictly specific for *G. max*? The microsyntenic arrangement of *hop* and chl-*fus* genes was determined for all 19 plant families under study (Figure 3 and species-specific details are shown in Additional file 1: Figure S1). In Chlorophyta, two families were mapped (Chlamydomonadaceae and Mamiellaceae) and each gene was found on a separate chromosome, suggesting the absence of microsynteny in this plant division. In return, 2 out of 3 studied families of monocots revealed the presence of *hop* and chl-*fus* genes on the same chromosome. Only in *Ensete ventricosum* (Musaceae), the pair of genes was found on separate chromosomes. In the same manner, the microsynteny is preserved in most of dicots excepting the Cucurbitaceae (3 species) and Fabaceae (3 out of 5 species) families, where the pair of genes is located on different chromosomes (Additional file: Table S1). In summary, the microsynteny of *hop* and chl-*fus* prevails in 78% (40 out of 51) of green plants studied. A graphic resume of microsynteny between *hop* and chl-*fus* genes among all plant species under study is shown in Additional file 3: Figure S2.

Concerning the one-to-one microcolinearity in convergent transcription of *hop* and chl-*fus*, three types of genome arrangements (*I* to *III*) were found in plants (Figure 4), as follows: *I*). Each gene resides on a different chromosome, i.e., they are not collinear (all Chlorophyta, one monocot, and six dicots). *II*) In Malvaceae (*Gossypium raimondii* and *Theobroma cacao*) the chl-*fus* gene moved just upstream of *hop* and both genes are transcribed in the same direction, i.e, local chromosome inversion [32,33]; and *III*) *hop* and chl-*fus* are colinear in convergent

transcription (no inserted elements), which is the most frequent arrangement in both monocots and dicots (34 out of 51 species analyzed or ≈ 66%).

Figure 3. Microsyntenic arrangement (at scale) of the pair of genes hop and chl-fus, among the 51 plant genomes under study. Hop protein TPR and DP domains are color-coded according to conventions (bottom boxes). IGR: intergenic region. Non-syntenic genes are drawn on separate chromosomes.



Interestingly, *Elaeis guineensis* and *Phoenix dactylifera* (monocots), as well as *Morus notabilis* and *Linum usitatissimum* (dicot) harbored sequences coding for retrovirus-like proteins within their intergenic sequences, i.e., inserted between *hop* and chl-*fus* genes (see the section about molecular instability of the intergenic region). Detailed physical maps for each species under study are shown in Additional file 1: Figure S1.

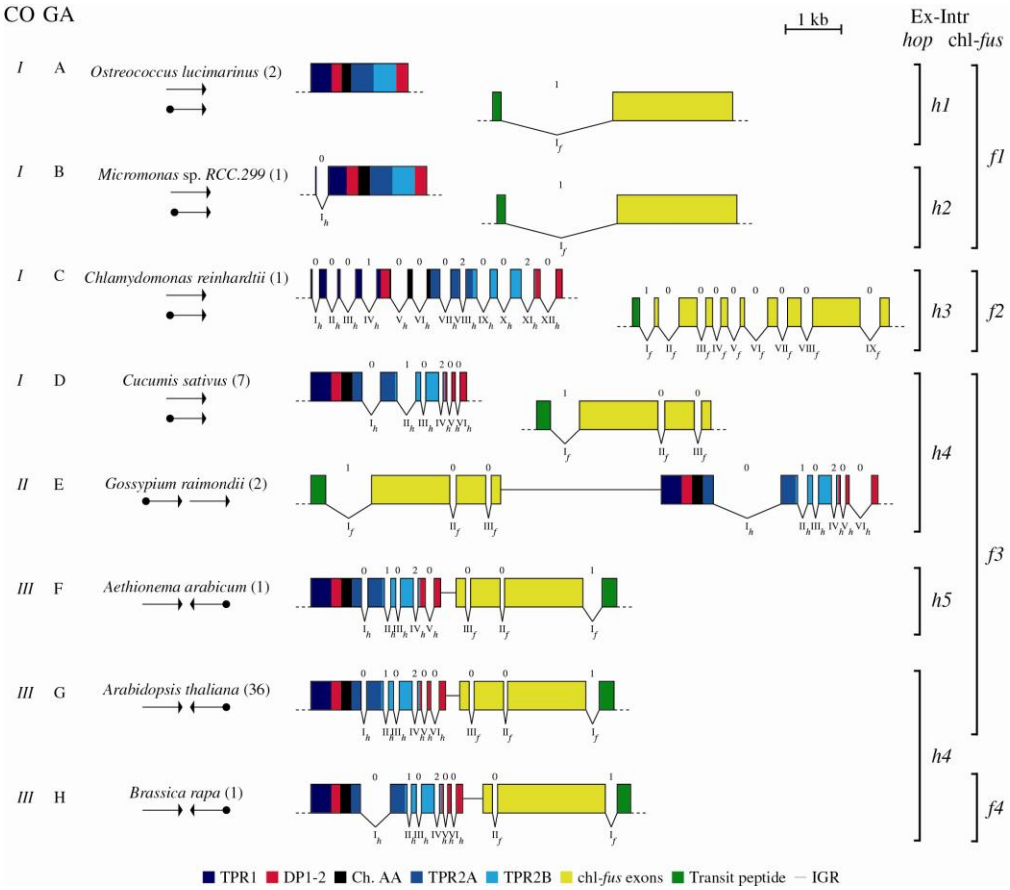
2.3. PARALLEL EVOLUTION OF EXON–INTRON GENE STRUCTURE OF *hop* AND chl-*fus* GENES

The human *hop* gene was found to contain 13 introns and intron phase was essential to hypothesize the evolutionary origin of Hop domains, by exon shuffling [6]. However, intron number and phase of plant *hop* genes are still unknown and this data could reinforce the role of introns in *hop* evolution from the initial stages of eukaryotic development. Therefore, we examined the exon–intron organization of *hop* and chl-*fus* genes among the 51 plant species, to infer the contribution of introns to the evolution of their resultant proteins (Table 1, Figures 3, 4 and Additional file 1: Figure S1).

The simultaneous spatial arrangement of exons and introns in the coding sequences of the pair *hop* − chl-*fus* in plants falls in one of eight categories (A to H), as shown in Figure 4. In type A (*O. lucimarinus* and *O. tauri*), *hop* lacks introns, while *fus* holds a single intron splitting the mature protein from the transit peptide-coding exon (labelled as I_f). Apparently, *Micromonas* sp. does not contain introns; however, it is very likely that a 5' intron is located after the first 18 nucleotides. An exceptionally long predicted Hop protein is registered in Genbank under the accession number XP_002500383; this polypeptide shares high identity with other plant Hop proteins, but contains 71 extra amino acids not found in any other eukaryote. A fine-scale analysis of this insertion suggests that an intron may have

gone unnoticed because it is in frame with a 5' short exon, coding for the conserved amino acids MADEHK. We show in Additional file 4: Figure S3 (A) an HCA alignment of predicted *Micromonas* sp. [GenBank:XP_002500383] and *A. thaliana* Hop proteins. In this alignment, a perfect match is evident between the two proteins excluding the extra 71 amino acids (bordered by a rounded rectangle).

Figure 4. Grouping of gene arrangements found for the pair of genes hop and chl-fus, among the 51 plant genomes under study. CO: classification by microcolinearity (categories *I* to *III*); GA: classification by gene arrangement, according to the exon–intron structure of both combined *hop* and chl-*fus* (categories A to H). Arabic numbers in parenthesis: number of species sharing the same gene arrangement; *hop* and chl-*fus* genes are represented by arrows to resume gene topology. Ex-Intr *hop*: exon–intron organizations found for *hop* gene (categories *h1* to *h5*), Ex-Intr chl-*fus*: exon–intron organizations found for chl-*fus* gene (categories *f1* to *f4*). Arabic and roman numbers represent intron phase (0, 1, or 2) and succession of introns from I to I+n, respectively; *hop* introns are named as I_h, II_h, III_h, etc., and chl-*fus* introns are named as I_f, II_f, III_f, etc. Exons coding for TPR and DP domains are color-coded according to conventions (bottom boxes). IGR: intergenic region. Non-syntenic genes are drawn on separate chromosomes.



In Additional file 4: Figure S3 (B), we represent the translated 5' regions of *Micromonas* sp. and predicted *C. reinhardtii hop* genes. We propose that nucleotides in bold belong to a phase-0 intron (I_h), which is in frame with the first and second exons. Conveniently, the exon–intron boundaries conserve the canonical splice consensus sequences AG:*GT* and *CAG*:GC [34,35]. According to this hypothesis, the predicted ORFs encode Hop proteins with the same number of amino acids than the other plant Hop members (Additional file 4: Figure S3 (C)). In addition, no significant similarity was found with a BLAST search using the 71 extra amino acids as query (not shown).

Taken together, these results led us to the conclusion that the *Micromonas* sp. *hop* gene must enclose one intron located just after the first seven codons (amino acids MADEHK). Thus, *Micromonas* sp. is classed in type B (Figure 4), in which both non-collinear genes have a single intron, i.e., 1-1 (Table I). In type C, (*C. reinhardtii*), *hop* contains 12 introns while *fus* has 9. Contrary to the other members of division Chlorophyta, *C. reinhardtii* has accumulated a noticeable plethora of introns; some of them lie in positions shared with human and higher plants (See next section). In type D (Musaceae (Monocot), Cucurbitaceae and 3 out of 5 Fabaceae (Dicot)) *hop* and chl-*fus* are not syntenic, but individual genes hold the same structure 6-3 of the greatest number of convergently transcribed genes in higher plants (type G). In type E, the exon–intron structure is the same of type G, but chl-*fus* was transposed to the 5' flanking site of *hop*, and transcribed in the same direction (Figure 4). In types F and H, *hop* and chl-*fus* lack one intron, respectively, with regard to type G. It is concluded that during the evolutionary process, *hop* and chl-*fus* genes underwent extensive changes in their exon−intron structure, among unicellular photosynthetic eukaryotes, as well as in higher plants. It is interesting to note that intron gain/loss affected both genes alike, by species. For example, *C. reinhardtii* (type C) *hop* and chl-*fus* conserved a plethora of introns (simultaneous intron gain?), while both genes in *O. lucimarinus* (type A) preserved

only one (simultaneous intron loss?). This finding also applies for higher plants (Figure 4).

2.4 INTRON POSITION AND PHASE AS DETERMINANT OF EXON SHUFFLING

In previous reports, it has been proposed that domain/module duplication has contributed to gene evolution through exon shuffling [36]. Bioinformatic analyses of vertebrate Hop orthologs suggested that TPR and DP domains behaved as a whole recombination unit due to the presence of phase-0 introns [6]. Phase-0 introns are the most favorable for exon duplication or shuffling without modifying the reading frame [36], and the human *hop* gene comprises TPR−DP modules bordered by phase-0 introns. Likewise, by primary structure sequence alignments, it was hypothesized that EF-G emerged as a result of gene duplication/fusion events [37].

We analyzed the exon–intron topologies and intron phase distribution within plant *hop* and chl-*fus* genes, in order to reconstruct the molecular events leading to the emergence of present-day genes. As shown in Figure 4, *hop* genes can be grouped in 5 classes of exon–intron structure (*h1-h5*), while *fus* genes are grouped in 4 (*f1-f4*). Considering only the *hop* gene, it contains one or more or none introns in green algae. No introns were found either in *Ostreococcus lucimarinus* or *Ostreococcus tauri* (Class *h1*), while *Micromonas* sp. was predicted to contain one 5’ phase-0 intron (Class *h2*). Contrary to the above mentioned Mamiellaceae family members (Figure 4), *C. reinhardtii* (Chlamydomonadaceae) is the photosynthetic eukaryote with the greatest number of introns, with 12 short intragenic regions equally distributed within the coding region (Class *h3*). Although most of introns are phase-0 (8 out of 12), the recombinable module that most resembles those found in vertebrates is located between phase-0 introns I_h to VI_h. This unit contains a complete TPR-DP-Ch. AA module, able to recombine by exon shuffling. Class *h4*

is the most abundant gene structure in higher plants (46 species). The first intron (I_h, phase-0) splits the TPR2A domain. The rest of introns (3 out of 5 of phase-0) split the end of the TPR2A-coding exons and the C-terminal TPR2B-DP2-coding sequences. Finally, Class *h5* (*Aethionema arabicum*, one member out of 9 of the Brassicaceae family) exhibits the same exon–intron topology of Class *h4*, except that it lacks the Class *h4* intron V_h, located within the DP2 domain (Figure 4).

Disparities in intron number were used to define classes *h1* to *h5* (Figure 4); Additional file 5: Figure S4 shows that not all intron positions are conserved among higher plants. For example, the first intron (phase-0) *in C. reinhardtii hop* gene (I_h), that locates between amino acids K and A, is also found in *Micromonas* sp. but not in either *O. lucimarinus*, *L. alabamica* or *A. arabicum*. The second intron (phase-0) *in C. reinhardtii* (II_h) locates between Y and A, and is exclusive to this species, and so forth. From Additional file 5: Figure S4 it is inferred that intron position is mainly conserved among *hop* genes from higher plants and partially between higher plants and Chlorophyta or plants and human. Specifically, the *C. reinhardtii* phase-0 I_h intron is shared with *Micromonas* sp. Likewise, *C. reinhardtii* introns II_h (0), III_h (0), IV_h (1), V_h (0), VI_h (0), $VIII_h$ (2), IX_h (0) and XI_h (2) are exclusive to this green alga. The *C. reinhardtii* VII_h (0) and XII_h (0) introns are shared with *L. alabamica* and *A. arabicum* and the rest of higher plants. Finally, higher plants contain introns restricted to Mono and Dicots, i.e., introns II_h (1), III_h (0) and IV_h (2). Exceptionally, *A. arabicum* (Brassicaceae, Class *h5*) lacks the phase-0 intron V_h of higher plants (Class *h4*). In the bottom of Additional file 5: Figure S4 we represent the human Hop protein and its related introns. A careful comparison of intron location among plants and human reveals that human Hop shares two introns with *C. reinhardtii* (i.e, I_h (0) and X_h (0)), but not with higher plants.

On the other hand, the chl-*fus* gene has undergone a higher reduction in intron number with respect to *hop*. The exon–intron structure was organized under four classes (*f1* to *f4*), according to the number and position of introns (Figure 4). From

algae to higher plants, the chl-*fus* gene contains a phase-1 intron that separates the signal peptide from the mature protein; this implies that a new exon coding for an N-terminal transit peptide was recruited, for the correct trafficking of cEF-G from cytoplasm to the plastids [38]. More precisely, Class *f1* embraces all predicted Mamiellaceae chl-*fus* genes with a single phase-1 intron, inserted between the chloroplast-targeting domain and the rest of the coding sequence (Figure 4). On the contrary, the *C. reinhardtii* (Chlamydomonadaceae) chl-*fus* gene has eight additional phase-0 introns interspersed within the cEF-G coding region (Class *f2*). Class *f3* is the most prevalent exon–intron organization found in monocot and dicot plants (Figure 4). It contains two phase-0 introns apart from that coding for the transit peptide (phase-1). Introns II_f and III_f are located within the 3' half of the chl-*fus* gene. Finally, only one member of Brassicaceae out of 9 (*Brassica rapa*) belongs to Class *f4*, which contains three exons and two introns. The *B. rapa* chl-*fus* gene lacks intron II_f with respect to Class *f3.*

2.5 MOLECULAR INSTABILITY OF THE *hop* AND chl-*fus* INTERGENIC REGION

In several plant families, the intergenic region (IGR) between the *hop* and chl-*fus* genes suffered insertions and deletions. While 74% of monocots and dicots preserve microcolinearity, the IGR among species is of variable length. For example, the shortest IGR belongs to *Leavenworthia alabamica* (188 bp), while the longest belong to *Linum usitatissimum* (38523 bp); nevertheless, the IGR region typically does not exceed 3500 bp (Additional file 1: Figure S1). IGR nucleotide sequences were analyzed by tBLASTn in order to identify potential ORFs. Plant retroviruses (or retrotransposons) and hypothetical genes were found in Monocots (*Elaeis guineensis* and *Phoenix dactylifera*) and Dicots (*Morus notabilis* and *Linum usitatissimum*), within IGRs >10 kb. For example, a putative pararetrovirus-like pseudogen was found within the 10 kb IGR of *M. notabilis*. In Additional file 6:

Figure S5 (A), we show a ClustalW alignment between a putative polyprotein encoded by the *M. notabilis* IGR and a *Citrus endogenous* pararetrovirus, retrieved by BLAST (45% identity). The *M. notabilis* predicted polyprotein is truncated by 12 aberrant stop codons, suggesting that it could be a pararetrovirus pseudogen. Furthermore, transposon-like repeated sequences were found in a number of species. For example, inverted repeat sequences of Miniature Inverted–Repeat Transposable Elements (MITEs) [39] were found within the IGR of *Oryza* spp (Additional file 6: Figure S5 (B)) and direct repeats of CACTA-like transposons reside in *M. truncatula* IGR (not shown).

Two interesting cases of deletions within the IGR have been found in higher plants, which alter the 3' untranslated region of the *hop* and chl-*fus* genes. In *Glycine max*, a plant with a predicted allopolyploidization event [40], two chl-*fus* genes were cloned and sequenced from cv. Ceresia (98% identity between cEF-G1 and cEF-G2 proteins), both with *hop* genes in convergent transcription [8]. ClustalW alignments were performed between chl-*fus* genes of *G. max* cv. Ceresia and cDNAs from *G. max* cv. Williams, which contain three different poly-A sites (Additional file 7: Figure S6 (A)). An almost perfect match was found between the coding part and the 3' untranslated region of the cDNAs, chl-*fus*1 and chl-*fus*2 genes; however, chl-*fus*1 drastically lacks identity 123 nucleotides downstream of the stop codon. A detailed nucleotide analysis allowed to conclude that a chromosomal deletion (ca. 680 bp) maps between the chl-*fus*1 and *hop*1 genes (Additional file 7: Figure S6 (B)).

A more severe case of IGR deletion is found in *A. thaliana*, in which the 3' transcribed regions of the *hop* and chl-*fus* genes overlap. We show in Additional file 8: Figure S7 a chromosomal map of the *A. thaliana hop* and chl-*fus* genes, and three cDNAs of each gene, with multiple poly-A sites. As can be observed, the 3' end of three *hop* and that of two chl-*fus* cDNAs overlap. Thus, in the strict sense, the IGR between *hop* and chl-*fus* genes is missing; nevertheless, according to the

Genbank cDNA accessions, both genes are transcribed. We concluded that the IGR separating the *hop* and chl-*fus* genes in plants seems to be a target region for insertion and deletion (indel) events, making it genetically unstable.

3. DISCUSSION

3.1 MICROSYNTENY AND COEVOLUTION OF *hop* AND chl-*fus* GENES IN PLANT GENOMES

In this report, we provided extensive evidences unveiling the evolutionary changes suffered by plant *hop* and chl-*fus* genes, after the primary endosymbiotic events. One gene is typically of nuclear origin, while the other came from the precursors of modern chloroplasts; together, they could constitute an interesting model to draw conclusions on the genome rearrangement events during and after the transfer of chloroplast genes to the nucleus. The first remark is the outstanding conservation of microsynteny and microcolinearity, in spite of all genomic duplications, deletions, inversions, insertions, and translocation events that shape genomes [1]. Nevertheless, our results in Figures 3 and 4 suggest that chl-*fus* was originally transferred from chloroplasts to a different chromosome from that of *hop* gene, in the proto-algal nuclear genome. This assumption is supported by the absence of microsynteny in green algae (prasinophytes), "which comprise the descendants of the primitive algae from which all green algal lineages, including the ancestors of land plants, evolved" [41,42]. Thus, the microcolinearity observed in higher plants should be the result of a recombination event, e.g., chromosome fusion, inversion or translocation [43], sometime before the appearance of monocots. A few monocot and dicot plant families also lack microsynteny, undoubtedly as a consequence of new genome rearrangements. While this issue rule out the possibility to discern details on the coevolution of nuclear vs. neighboring laterally transferred genes, each gene provides new insights to reconstruct the history of ancient nuclear genes.

The second interesting finding is the high degree of conservation of their encoded proteins, across evolution. Both genes become from domain or module duplications [6,36,37,44] but these events happened very early in time, before

further intron gain and losses [45]. The phylogenetic trees in Figures 1 and 2 reveal a high conservation of Hop and cEF-G proteins, in opposition to gene structure (Figures 3 and 4) and DNA sequences (not shown), indicating that the conservation of their 1D to 3D structures are essential for their cellular functions. In all photosynthetic organisms under study, Hop keeps the typical domain structure of the fungi and animal orthologs (Additional file 5: Figure S4) [6]. This is an unexpected finding because in fungi, nematodes or insects, isoforms of the Hop protein lack DP1 or TPR1-DP1 domains [46], and it was hypothesized the existence of deletion mutants in plants; therefore, the DP1-mutant found in *G. max* [6] is actually an exception rather than the rule. On the other hand, the cEF-G protein also remained virtually unchanged with respect to its prokaryotic ancestor (Figure 1). Although plant cEF-G exhibits higher similarity with bacterial EF-G proteins, it shows a closer phylogenetic relationship with α-proteobacteria rather than with cyanobacteria, suggesting that the ancestor of cEF-G could be the α-proteobacterial progenitor of mitochondria [20]. Our results, based on the analysis of 51 plant species of 19 families, support that hypothesis without exception. Furthermore, it has been reported that two isoforms of EF-G have distinct roles in both translocation (EF-G1) and ribosome recycling (EF-G2) in a variety of species from bacteria [47] to mammals [48]. Phylogenetic trees built with a few of plant cEF-G sequences evidenced that cEF-G does not fall within one of these categories and forms a separate clade [20]; our phylogenetic analysis of 51 cEF-G sequences confirm this finding and demonstrate the existence of a single form of cEF-G proteins in photosynthetic organisms (Figure 1). Thus, chloroplast protein synthesis translocation and ribosome recycling functions might be assumed by that unique form of cEF-G.

3.2 ROLE OF INTRONS IN *hop* GENE EVOLUTION

The observed exon-intron structure of *hop* and chl-*fus* at different levels of organismal complexity (Figures 3 and 4) leads to three main conclusions: Firstly,

evidences support the hypothesis that both genes experienced intron gain and losses, before and after the transfer of chl-*fus* to the nuclear genome (Figure 4). Secondly, whenever a gene gained (or lost) introns, the other did too, suggesting a species-specific synchronized intron gain/loss: for example, in *Micromonas* sp. both non-collinear genes have a single intron, but in *C. reinhardtii* they gained multiple introns each [49,50]. Finally, exon shuffling played essential roles in the construction of these genes, making it feasible to reconstruct their evolutionary changes. Inexorably, recombination of symmetric exons/modules would keep the open reading frame uninterrupted by frameshifts [51-53].

It has been proposed that in vertebrates, the *hop* gene could have emerged from recombinable modules surrounded by introns of the same phase [6]. Our results provide new evidences that phase-0 introns were essential for *hop* gene construction in all eukaryotes. Based on the five gene topologies of Figure 4 (*h1* to *h5*), we propose a model of the ancient events giving rise to the present-day structure of *hop* genes, with a minimum number of steps (see Additional file 9: Figure S8 and legend). Our model leads to some significant conclusions of the role of introns in *hop* gene evolution: i) Phase-0 introns were critical for serial exon shuffling recombinations of a primordial module [25,36,53,54] composed of symmetric exons «miniexon – phase-0 intron – TPR domain − phase-0 intron − DP domain − phase-0 intron –Ch. AA domain », and giving rise to a 'Proto-eukaryote *hop'*. Old phase-0 introns could be traced backward in time (i.e., green and purple, Additional file 9: Figure S8), a typical characteristic of ancient proteins constructed by shuffling of exon/modules [25,45,55,56]. According to our evolutionary model, the human *hop* would preserve two old phase-0 introns (I_h and V_h) as reminiscent of the original recombinable modules. ii) The origin of introns is still a matter of debate [44,57-59]. Nevertheless, it is difficult to explain the differences in intron number and position within *hop* genes, between animals and plants for example, or between *C. reinhardtii* and *Micromonas* sp., without considering a recent gain/loss of introns. According to our model, the gain/loss of introns by *hop* was a very

dynamic process, leading to conclude that while some (phase-0) introns are very old, other ones (phase-0, 1 and 2) might be of recent origin, a long-standing hypothesis proposed for other eukaryotic genes e.g., the triose-phosphate isomerase gene [60]. Nevertheless, even though the gene was subjected to many recombinations, the ORF remained virtually unchanged (Figure 2), excepting some shorter isoforms of undiscovered functions [6]. iii) It has been noticed a biased distribution of phase-0 introns immediately after the start codon in eukaryotic genes (vertebrates, invertebrates, fungi, plants, and protists), specially "at the boundaries of evolutionary modules in proteins without signal peptides and this effect is stronger in phylogenetically old proteins" [45,61,62]. Authors suggest that these introns should "allow the 5' UTR to participate in exon shuffling, so that different genes can exchange regulatory information" [62]. Interestingly, some present-day *hop* genes (Micromonas, *C. reinhardtii*, human) share a phase-0 intron (i.e., the green intron of Additional file 9: Figure S8) downstream of the first 3 to 6 amino acids (Figure 4 and Additional file 5: Figure S4). Since *hop* genes are regulated by different forms of stress [8,24,63], it is conceivable that it was by this way that the gene became stress-regulated. However, this well-disposed intron also could contribute to shuffle internal exons, specifically whole TPR or TPR-DP domains, a valid assumption in support of our evolutionary model (Additional file 9: Figure S8).

3.3 ROLE OF INTRONS IN chl-*fus* GENE EVOLUTION

It has been proposed that the *fus* gene is actually a product of three consecutive duplication/fusion gene events [37]. Such kind of successive duplication/fusions of peptide segments becomes conceivable with the presence of phase-0 introns. However, since chloroplasts, and then *fus* genes are of prokaryotic origin, probably introns had not a significant role in the creation of the primordial *fus*, but some kind of illegitimate recombination [64]. Thus, present-day spliceosomal introns (all phase-0) very likely were gained after the transfer of chloroplast DNA to the nucleus [57]. Nevertheless, the phase-1 intron connecting the N-terminal transit

peptide-coding exon and the mature protein may have played an important role in the functional establishment of chl-*fus* in the nucleus (Figure 4) and its loss from the chloroplast. Certainly, experimental evidence supports the assumption that chloroplasts transfer genes to the nucleus at high frequencies; however, the rate of nuclear establishment is extremely low. This conclusion is supported by the low number of loci encoding transferred genes [9,65-67]. Interestingly, all the 51 chl-*fus* genes under study contain this intron (Additional file 1: Figure S1). How intron I_f was acquired? Exon shuffling has been proposed as the main form for the gain of transit peptide exons. Moreover, the gain of intron I_f must have included the transit peptide-coding exon and probably regulatory sequences [68].

Why phase-1 and not 0 or 2? A recent study on human secretory signal peptides revealed a biased distribution of phase-1 introns (49,9%), in the vicinity of the signal peptide cleavage sites [69]. According to the authors “phase-1 introns most frequently split the four G↓GN codons encoding glycine”, that “are significantly enriched in positions -1, -3, -4 and -5”. Instead of this, for chl-*fus* genes, virtually all monocot and dicot phase-1 introns split codons G↓AU or G↓AC (Asp), and G↓AA or G↓AG (Glu), all fairly frequent split codons in all eukaryote taxonomic groups [70]. Exceptionally, *B. distachyon* (G↓GT) contains a triplet coding for the widespread Gly. Interestingly, this exception also applies for Chlorophyta: While *C. reinhardtii* keeps a G↓AC codon (Asp), *Micromonas*, *O. lucimarinus* and *O. tauri* contain G↓CN (Ala). Thus, it is tempting to speculate that the phase-1 intron that favored the fusion with the transit peptide-coding exon was originally splitting a G↓CN codon (Ala). Sometime in the evolution before the appearance of *C. reinhardtii*, G↓CN mutated to G↓AN (a C to A transversion).

3.4 WOULD BE COMPROMISED THE INTEGRITY OF THE chl-*fus* GENE FOR THE FUTURE?

It has been elucidated that *hop* genes have a long history of gene rearrangements, which ended in the present-day form. These evidences support a natural susceptibility of the intergenic region to recombine: i) The chl-*fus* gene was recombined downstream of *hop* and this interchanging might not have been a coincidence. ii) The IGR between *hop* and chl-*fus* has been in the midst of new chromosome rearrangements (e.g., gene inversion in Malvaceae); such events must require some molecular propensity of that DNA to recombine. iii) We showed that in some plant species, retroviruses found suitable nucleotide sequences for transposition within the IGR. iv) Strikingly, in *G. max* the IGR almost disappeared, and in *A. thaliana*, it is totally absent. Thus, the unavoidable question is, where that propensity to recombine comes from? In our sequence analyses, we found a wide set of mobile elements inserted within the IGR of both monocots and dicots, revealing its proclivity to recombine. Interestingly, CACTA elements “frequently transduce host sequences” [71]; thus the presence of mobile DNA reinforces our assumption of a site of chromosomal instability. Currently, there is no database available for an extensive search of recombination "hot spots" [72], covering all the plant species studied here. However the possibility that chl-*fus* and *hop* genes are in the middle of a recombination "hot spot" should not be discarded. Regardless of the basis of such DNA instability, it is that the tendency to gain or loss nucleotides has come to affect the integrity of 3’ flanking sequences. Since there are no other genetic loci coding for the cEF-G protein (contrary to *hop* gene families), there would be a real risk of having plant mutants lacking the whole or part of the chl-*fus* genes. Actually, it may already have happened a number of times but such mutants could be unviable, in theory. Paradoxically, the chl-*fus* gene was transposed into a point of DNA instability and heretofore it continues to occupy the same and unique locus in the plant genome, judging by the high conserved micro-synteny.

4. CONCLUSIONS

In this study, we performed a deep analysis of the structure of two convergently transcribed nuclear genes, *hop* (nuclear origin) and chl-*fus* (plastid origin). We concluded that their convergence was a product of chromosome recombination rather than direct transfer of chl-*fus* from the chloroplast, downstream of *hop*. The exon–intron organization and intron phase of both genes agree with exon shuffling events, giving rise to exon/module duplications and transit peptide recruiting for chloroplast protein import. We showed evidences of instability of the intergenic region and susceptibility to recombination, that could favored the recombination of chl-*fus* within this region. Finally, the pair of genes *hop* and chl-*fus* should be useful as genetic markers, on the basis of microcolinearity in higher plants but not in Chlorophyta.

REFERENCES

1. Tang H, Bowers JE, Xiyin W, Ming R, Alam M, Paterson AH: **Synteny and colinearity in plant genomes**. *Science* 2008, **320:**486−488.

2. McCouch SR: **Genomics and Synteny.** *Plant Physiol* 2001, **125:**152−155.

3. Johnson BD, Schumacher RJ, Ross ED, Toft DO: **Hop modulates Hsp70/Hsp90 interactions in protein folding.** *J Biol Chem* 1998, **273:**3679−3686.

4. Chen S, Smith DF: **Hop as an adaptor in the heat shock protein 70 (Hsp70) and hsp90 chaperone machinery.** *J Biol Chem* 1998, **273:**35194−35200.

5. Scheufler C, Brinker A, Bourenkov G, Pegorano S, Moroder L, Bartunik H, Harti FU, moarefi I: **Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine.** *Cell* 2000, **101:**199−210.

6. Hernández Torres J, Papandreou N, Chomilier J: **Sequence analyses reveal that a TPR-DP module, surrounded by recombinable flanking introns, could be at the origin of eukaryotic Hop and Hip TPR-DP domains and prokaryotic GerD proteins.** *Cell Stress Chaperones* 2009, **14:**281−289.

7. Odunuga OO, Longshaw VM, Blatch GL: **Hop: more than an Hsp70/Hsp90 adaptor protein.** *Bioessays* 2004, **26:**1058−1068.

8. Hernández Torres J, Chatellard P, Stutz E: **Isolation and characterization of *gmsti*, a stress-inducible gene from soybean (*Glycine max*) coding for a**

protein belonging to the TPR (tetratricopeptide repeats) family. *Plant Mol Biol* 1995, **27:**1221−1226.

9. Martin W, Herrmann RG: **Gene transfer from organelles to the nucleus: how much, what happens, and Why?** *Plant Physiol* 1998, **118:**9−17.

10. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D: **Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.** *Proc Natl Acad Sci USA* 2002, **99:**12246−12251.

11. Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T: **Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor.** *Mol Biol Evol* 2008, **25:**748−761.

12. Chua NH, Schmidt GW: **Transport of proteins into mitochondria and chloroplasts.** *J Cell Biol* 1979, **81:**461−483.

13. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M; **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression.** *EMBO J* 1986, **5:**2043−2049.

14. Ohyama K, Fukuzawa H, Kohchi T, Sano T, Sano S, Shirai H, Umesono K, Shiki Y, Takeuchi M, Aota Z, Inokuchi H, Ozeki H: **Structure and organization of *Marchantia polymorpha* chloroplast genome: I. Cloning and gene identification.** *J Mol Biol* 1988, **203:**281−298.

15. Hiratsuka J,Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M: **The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals.** *Mol Gen Genet* 1989, **217:**185−194.

16. Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E: **Complete sequence of *Euglena gracilis* chloroplast DNA.** *Nucleic Acids Res* 1993, **21:**3537−3544.

17. Breitenberger CA, Spremulli LL: **Purification of *Euglena gracilis* chloroplast elongation factor G and comparison with other prokaryotic and eukaryotic translocases.** *J Biol Chem* 1980, **255:**9814−9820.

18. Hernández Torres J, Breitenberger CA, Spielmann A, Stutz E: **Cloning and sequencing of a soybean nuclear gene coding for a chloroplast translation elongation factor EF-G.** *Biochim Biophys Acta* 1993, **1174:**191−194.

19. Girshovich AS, Kurtskhalia TV, Ovchinnikov YA, Vasiliev, VD: **Localization of the elongation factor G on *Escherichia coli* ribosome.** *FEBS Lett* 1981, **130:**54−59.

20. Atkinson GC, Baldauf SL: **Evolution of elongation factor G and the origins of Mitochondrial and Chloroplast Forms.** *Mol Biol Evol* 2011, **28:**1281−1292.

21. Wilson KS, Noller HF: **Molecular movement inside the translational engine.** *Cell* 1998, **92:**337−349.

22. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2013, **41:**36−42.

23. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A: **Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.** *Bioinformatics* 2012, **28:**1647−1649.

24. Honoré B, Leffers H, Madsen P, Rasmussen HH, Vanderkerchhove J, Celis JE: **Molecular cloning and expression of a transformation-sensitive human protein containing the TPR motif and sharing identity to the stress-inducible yeast protein STI1.** *J Biol Chem* 1992, **267:**8485−8491.

25. Patthy L: **Intron-dependent evolution: Preferred types of exons and introns.** *FEBS Lett* 1987, **214:**1−7.

26. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688−2690.

27. Callebaut I, Labesse G, Drand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP: **Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives.** *Cell Mol Life Sci* 1997, **53:**621−645.

28. Woodcock S, Mornon JP, Henrissat B: **Detection of secondary structure elements in proteins by hydrophobic cluster analysis.** *Protein* Eng 1992, **5:**629−635.

29. Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I: **A generalized analysis of hydrophobic and loop clusters within globular protein sequences.** *BMC Struct Biol* 2007, **7:**2.

30. Akkaya MS, Welcsh PL, Wolfe MA, Duerr BK, Becktel WJ, Breitenberger CA: **Purification and N-terminal sequence analysis of pea chloroplast protein synthesis factor EF-G.** *Arch Biochem Biophys* 1994, **308:**109−117.

31. Felsenstein J: **Confidence limits on phylogenies: An approach using the bootstrap.** *Evolution* 1985, **39:**783−791.

32. Kirkpatrick M, Barton N: **Chromosome inversions, local adaptation and speciation.** *Genetics* 2006, **173:**419−434.

33. Kirkpatrick M: **How and why chromosome inversions evolve.** *PLoS Biol* 2010, **8:**e1000501.

34. Qiu WG, Schisler N, Stoltzfus A: **The evolutionary gain of spliceosomal introns: Sequence and phase preferences.** *Mol Biol Evol* 2004, **21:**1252−1263.

35. Carmel L, Rogozin IB, Wolf YI, Koonin EV: **Patterns of intron gain and conservation in eukaryotic genes.** *BMC Evol Biol* 2007, **7:**192.

36. Patthy L: **Genome evolution and the evolution of exon-shuffling--a review.** *Gene* 1999, **238:**103−114.

37. Cousineau B, Leclerc F, Cedergren R: **On the origin of protein synthesis factors: a gene duplication/fusion model.** *J Mol Evol* 1997, **45:**661−670.

38. Timmis JN, Ayliffe MA, Huang CY, Martin W: **Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes.** *Nature Rev Genet* 2004, **5:**123−135.

39. Momose M, Abe Y, Ozeki Y: **Miniature inverted-repeat transposable elements of Stowaway are active in potato.** *Genetics* 2010, **186:**59−66.

40. Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA: **Molecular and chromosomal evidence for allopolyploidy in soybean.** *Plant Physiol* 2009, **151:**1167−1174.

41. Nedelcu AM, Miles lH, Fagir AM, Karol K: **Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals.** *J Evol Biol* 2008, **21:**1852−1860.

42. Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O: **Phylogeny and molecular evolution of the green algae.** *Crit Rev Plant Sci* 2012, **31:**1−46.

43. Schubert I: **Chromosome evolution.** *Curr Opin Plant Biol* 2007, **10:**109−115.

44. Long M, De Souza SJ, Gilbert W: **Evolution of the intron-exon structure of eukaryotic genes.** *Curr Opin Genetics Dev* 1995, **5:**774−778.

45. De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: **Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins.** *Proc Natl Acad Sci U.S.A.* 1998, **95:**5094−5099.

46. Flom G, Behal RH, Rosen L, Cole DG, Johnson JL: **Definition of the minimal fragments of Sti1 required for dimerization, interaction with Hsp70 and Hsp90 and in vivo functions.** *Biochem J* 2007, **404:**159−167.

47. Suematsu T, Yokobori S, Morita H, Yoshinari S, Ueda T, Kita K, Takeuchi N, Watanabe Y: **A bacterial elongation factor G homologue exclusively functions in ribosome recycling in the spirochaete *Borrelia burgdorferi*.** *Mol Microbiol* 2010, **75:**1445−1454.

48. Tsuboi M, Morita H, Nozaki Y, Akama K, Ueda T, Ito K, Nierhaus KH, Takeuchi N: **EF-G2mt is an exclusive recycling factor in mammalian mitochondrial protein synthesis.** *Mol Cell* 2009, **35:**502−510.

49. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in Eukaryotic evolution.** *Curr Biol* 2003, **13:**1512−1517.

50. Roy SW, Penny D: **Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution.** *Genome Res* 2006, **16:**1270−1275.

51. Kolkman JA, Stemmer WPC: **Directed evolution of proteins by exon shuffling.** *Nat Biotechnol* 2001, **19:**423−428.

52. Jia Y, Keong C: **Statistical analysis of symmetric exon sets in eukaryotic genes.** *Genome Inform* 2003, **14:**410−411.

53. Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD: **Can codon usage bias explain intron phase distributions and exon symmetry?** *J Mol Evol* 2005, **60:**99−104.

54. França GS, Cancherini DV, De Souza SJ: **Evolutionary history of exon shuffling.** *Genetics* 2012, **140:**249−257.

55. De Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W: **Intron positions correlate with module boundaries in ancient proteins.** *Proc Natl Acad Sci USA* 1996, **93:**14632−14636.

56. Björklund AK, Ekman D, Elofsson A: **Expansion of protein domain repeats.** *PLoS Comput Biol* 2006, **2:**e114.

57. Jeffares DC, Mourier T, Penny D: **The biology of intron gain and loss.** *Trends Genet* 2006, **22:**16−22.

58. Rodríguez-Trelles F, Tarrio R, Ayala FJ: **Origins and evolution of spliceosomal introns.** *Annu Rev Genet* 2006, **40:**47−76.

59. Penny D, Hoeppner MP, Poole AM, Jeffares DC: **An overview of the introns-first theory.** *J Mol Evol* 2009, **69:**527−540.

60. Logsdon JM Jr, Tyshenko MG, Dixon C, Jarafi J, Walker VK, Palmer JD: **Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory.** *Proc Natl Acad Sci USA* 1995, **92:**8507−8511.

61. Fedorov A, Roy S, Cao X, Gilbert W: **Phylogenetically older introns strongly correlate with module boundaries in ancient proteins.** *Genome Res* 2003, **13:**1155−1157.

62. Nielsen H, Wernersson R: **An overabundance of phase-0 introns immediately after the start codon in eukaryotic genes.** *BMC Genomics* 2006, **7:**256.

63. Nicolet CM, Craig EA: **Isolation and characterization of STI1, a stress-inducible gene from *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1989, **9**:3638−3646.

64. Van Rijk A, Bloemendal H: **Molecular mechanisms of exon shuffling: illegitimate recombination.** *Genetics* 2003, **118:**245−249.

65. Huang CY, Ayliffe MA, Timmis JN: **Direct measurement of the transfer rate of chloroplast DNA into the nucleus.** *Nature* 2003, **422:**72−76.

66. Sheppard AE, Timmis JN: **Instability of plastid DNA in the nuclear genome.** *PLoS Genet* 2009, **5:**e1000323.

67. Stegemann S, Hartmann S, Ruf S, Bock R: **High-frequency gene transfer from the chloroplast genome to the nucleus.** *Proc Natl Acad Sci USA* 2003, **100:**8828−8833.

68. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD: **Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron.** *EMBO J* 1991, **10:**3073−3078.

69. Tordai H, Patthy L: **Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides.** *FEBS Lett* 2004, **575:**109-111.

70. Tomita M, Shimizu N, Brutlag DL: **Introns and reading frames: correlation between splicing sites and their codon positions.** *Mol Biol Evol* 1996, **13:**1219−1223.

71. Feschotte C, Pritham EJ: **DNA transposons and the evolution of eukaryotic genomes.** *Annu Rev Genet* 2007, **41:**331−368.

72. Mézard C: **Meiotic recombination hotspots in plants.** *Biochem Soc Trans* 2006, **34:**531−534.

73. Ouyang S, Buell CR: **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants.** *Nucleic Acids Res* 2004, **32:**360−363.

74. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31:**3406−3415.

BIBLIOGRAPHY

Akkaya MS, Welcsh PL, Wolfe MA, Duerr BK, Becktel WJ, Breitenberger CA: Purification and N-terminal sequence analysis of pea chloroplast protein synthesis factor EF-G. *Arch Biochem Biophys* 1994, 308:109−117.

Atkinson GC, Baldauf SL: Evolution of elongation factor G and the origins of Mitochondrial and Chloroplast Forms. *Mol Biol Evol* 2011, 28:1281−1292.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: GenBank. *Nucleic Acids Res* 2013, 41:36−42.

Björklund AK, Ekman D, Elofsson A: Expansion of protein domain repeats. *PLoS Comput Biol* 2006, 2:e114.

Breitenberger CA, Spremulli LL: Purification of *Euglena gracilis* chloroplast elongation factor G and comparison with other prokaryotic and eukaryotic translocases. *J Biol Chem* 1980, 255:9814−9820.

Callebaut I, Labesse G, Drand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP: Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* 1997, 53:621−645.

Carmel L, Rogozin IB, Wolf YI, Koonin EV: Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol* 2007, 7:192.

Chen S, Smith DF: Hop as an adaptor in the heat shock protein 70 (Hsp70) and hsp90 chaperone machinery. *J Biol Chem* 1998, 273:35194−35200.

Chua NH, Schmidt GW: Transport of proteins into mitochondria and chloroplasts. *J Cell Biol* 1979, 81:461−483.

Cousineau B, Leclerc F, Cedergren R: On the origin of protein synthesis factors: a gene duplication/fusion model. *J Mol Evol* 1997, 45:661−670.

De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci U.S.A.* 1998, 95:5094−5099.

De Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W: Intron positions correlate with module boundaries in ancient proteins. *Proc Natl Acad Sci USA* 1996, 93:14632−14636.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T: Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol* 2008, 25:748−761.

Eudes R, Le Tuan K, Delettre J, Mornon JP, Callebaut I: A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct Biol* 2007, 7:2.

Fedorov A, Roy S, Cao X, Gilbert W: Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res* 2003, 13:1155−1157.

Felsenstein J: Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 1985, 39:783−791.

Feschotte C, Pritham EJ: DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 2007, 41:331−368.

Flom G, Behal RH, Rosen L, Cole DG, Johnson JL: Definition of the minimal fragments of Sti1 required for dimerization, interaction with Hsp70 and Hsp90 and in vivo functions. *Biochem J* 2007, 404:159−167.

França GS, Cancherini DV, De Souza SJ: Evolutionary history of exon shuffling. *Genetics* 2012, 140:249−257.

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD: Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J* 1991, 10:3073−3078.

Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA: Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol* 2009, 151:1167−1174.

Girshovich AS, Kurtskhalia TV, Ovchinnikov YA, Vasiliev, VD: Localization of the elongation factor G on *Escherichia coli* ribosome. *FEBS Lett* 1981, 130:54−59.

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E: Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res* 1993, 21:3537−3544.

Hernández Torres J, Breitenberger CA, Spielmann A, Stutz E: Cloning and sequencing of a soybean nuclear gene coding for a chloroplast translation elongation factor EF-G. *Biochim Biophys Acta* 1993, 1174:191−194.

Hernández Torres J, Chatellard P, Stutz E: Isolation and characterization of *gmsti*, a stress-inducible gene from soybean (*Glycine max*) coding for a protein belonging to the TPR (tetratricopeptide repeats) family. *Plant Mol Biol* 1995, 27:1221−1226.

Hernández Torres J, Papandreou N, Chomilier J: Sequence analyses reveal that a TPR-DP module, surrounded by recombinable flanking introns, could be at the origin of eukaryotic Hop and Hip TPR-DP domains and prokaryotic GerD proteins. *Cell Stress Chaperones* 2009, 14:281−289.

Hiratsuka J,Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR, Meng BY, Li YQ, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M: The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol Gen Genet* 1989, 217:185−194.

Honoré B, Leffers H, Madsen P, Rasmussen HH, Vanderkerchhove J, Celis JE: Molecular cloning and expression of a transformation-sensitive human protein containing the TPR motif and sharing identity to the stress-inducible yeast protein STI1. *J Biol Chem* 1992, 267:8485−8491.

Huang CY, Ayliffe MA, Timmis JN: Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* 2003, 422:72−76.

Jeffares DC, Mourier T, Penny D: The biology of intron gain and loss. *Trends Genet* 2006, 22:16−22.

Jia Y, Keong C: Statistical analysis of symmetric exon sets in eukaryotic genes. *Genome Inform* 2003, 14:410−411.

Johnson BD, Schumacher RJ, Ross ED, Toft DO: Hop modulates Hsp70/Hsp90 interactions in protein folding. *J Biol Chem* 1998, 273:3679−3686.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A: Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012, 28:1647−1649.

Kirkpatrick M, Barton N: Chromosome inversions, local adaptation and speciation. *Genetics* 2006, 173:419−434.

Kirkpatrick M: How and why chromosome inversions evolve. *PLoS Biol* 2010, 8:e1000501.

Kolkman JA, Stemmer WPC: Directed evolution of proteins by exon shuffling. *Nat Biotechnol* 2001, 19:423−428.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O: Phylogeny and molecular evolution of the green algae. *Crit Rev Plant Sci* 2012, 31:1−46.

Logsdon JM Jr, Tyshenko MG, Dixon C, Jarafi J, Walker VK, Palmer JD: Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci USA* 1995, 92:8507−8511.

Long M, De Souza SJ, Gilbert W: Evolution of the intron-exon structure of eukaryotic genes. *Curr Opin Genetics Dev* 1995, 5:774−778.

Martin W, Herrmann RG: Gene transfer from organelles to the nucleus: how much, what happens, and Why? *Plant Physiol* 1998, 118:9−17.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D: Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 2002, 99:12246−12251.

McCouch SR: Genomics and Synteny. *Plant Physiol* 2001, 125:152−155.

Mézard C: Meiotic recombination hotspots in plants. *Biochem Soc Trans* 2006, 34:531−534.

Momose M, Abe Y, Ozeki Y: Miniature inverted-repeat transposable elements of Stowaway are active in potato. *Genetics* 2010, 186:59−66.

Nedelcu AM, Miles lH, Fagir AM, Karol K: Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals. *J Evol Biol* 2008, 21:1852−1860.

Nicolet CM, Craig EA: Isolation and characterization of STI1, a stress-inducible gene from *Saccharomyces cerevisiae*. *Mol Cell Biol* 1989, 9:3638−3646.

Nielsen H, Wernersson R: An overabundance of phase-0 introns immediately after the start codon in eukaryotic genes. *BMC Genomics* 2006, 7:256.

Odunuga OO, Longshaw VM, Blatch GL: Hop: more than an Hsp70/Hsp90 adaptor protein. *Bioessays* 2004, 26:1058−1068.

Ohyama K, Fukuzawa H, Kohchi T, Sano T, Sano S, Shirai H, Umesono K, Shiki Y, Takeuchi M, Aota Z, Inokuchi H, Ozeki H: Structure and organization of *Marchantia polymorpha* chloroplast genome: I. Cloning and gene identification. *J Mol Biol* 1988, 203:281−298.

Ouyang S, Buell CR: The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* 2004, 32:360−363.

Patthy L: Genome evolution and the evolution of exon-shuffling--a review. *Gene* 1999, 238:103−114.

Patthy L: Intron-dependent evolution: Preferred types of exons and introns. *FEBS Lett* 1987, 214:1−7.

Penny D, Hoeppner MP, Poole AM, Jeffares DC: An overview of the introns-first theory. *J Mol Evol* 2009, 69:527−540.

Qiu WG, Schisler N, Stoltzfus A: The evolutionary gain of spliceosomal introns: Sequence and phase preferences. *Mol Biol Evol* 2004, 21:1252−1263.

Rodríguez-Trelles F, Tarrio R, Ayala FJ: Origins and evolution of spliceosomal introns. *Annu Rev Genet* 2006, 40:47−76.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in Eukaryotic evolution. *Curr Biol* 2003, 13:1512−1517.

Roy SW, Penny D: Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. *Genome Res* 2006, 16:1270−1275.

Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD: Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol* 2005, 60:99−104.

Scheufler C, Brinker A, Bourenkov G, Pegorano S, Moroder L, Bartunik H, Harti FU, moarefi I: Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* 2000, 101:199−210.

Schubert I: Chromosome evolution. *Curr Opin Plant Biol* 2007, 10:109−115.

Sheppard AE, Timmis JN: Instability of plastid DNA in the nuclear genome. *PLoS Genet* 2009, 5:e1000323.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M; The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* 1986, 5:2043−2049.

Stamatakis A: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006, 22:2688−2690.

Stegemann S, Hartmann S, Ruf S, Bock R: High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc Natl Acad Sci USA* 2003, 100:8828−8833.

Suematsu T, Yokobori S, Morita H, Yoshinari S, Ueda T, Kita K, Takeuchi N, Watanabe Y: A bacterial elongation factor G homologue exclusively functions in ribosome recycling in the spirochaete *Borrelia burgdorferi*. *Mol Microbiol* 2010, 75:1445−1454.

Tang H, Bowers JE, Xiyin W, Ming R, Alam M, Paterson AH: Synteny and colinearity in plant genomes. *Science* 2008, 320:486−488.

Timmis JN, Ayliffe MA, Huang CY, Martin W: Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Rev Genet* 2004, 5:123−135.

Tomita M, Shimizu N, Brutlag DL: Introns and reading frames: correlation between splicing sites and their codon positions. *Mol Biol Evol* 1996, 13:1219−1223.

Tordai H, Patthy L: Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. *FEBS Lett* 2004, 575:109-111.

Tsuboi M, Morita H, Nozaki Y, Akama K, Ueda T, Ito K, Nierhaus KH, Takeuchi N: EF-G2mt is an exclusive recycling factor in mammalian mitochondrial protein synthesis. *Mol Cell* 2009, 35:502−510.

Van Rijk A, Bloemendal H: Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetics* 2003, 118:245−249.

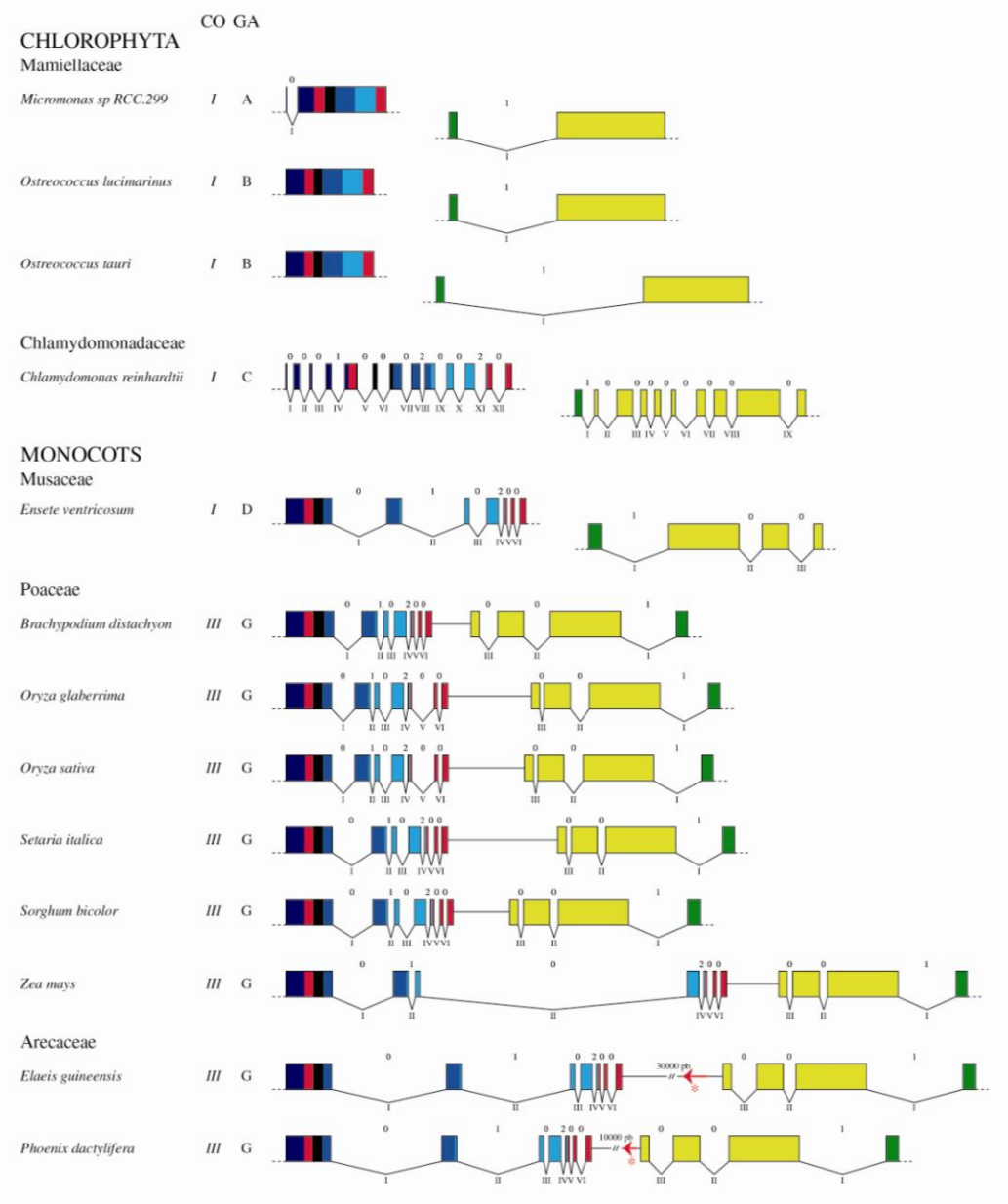
Wilson KS, Noller HF: Molecular movement inside the translational engine. *Cell* 1998, 92:337−349.

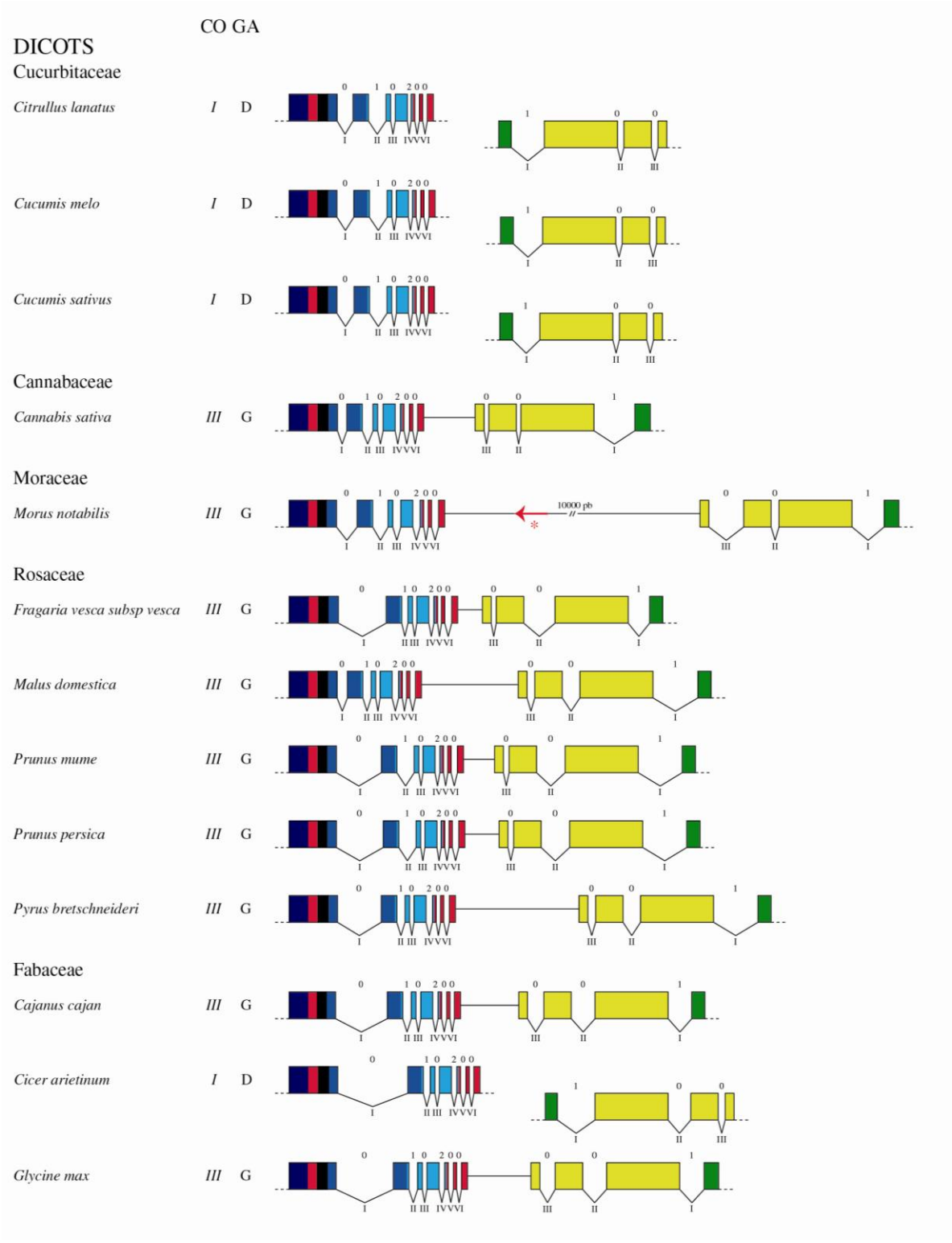
Woodcock S, Mornon JP, Henrissat B: Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein* Eng 1992, 5:629−635.
Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003, 31:3406−3415.

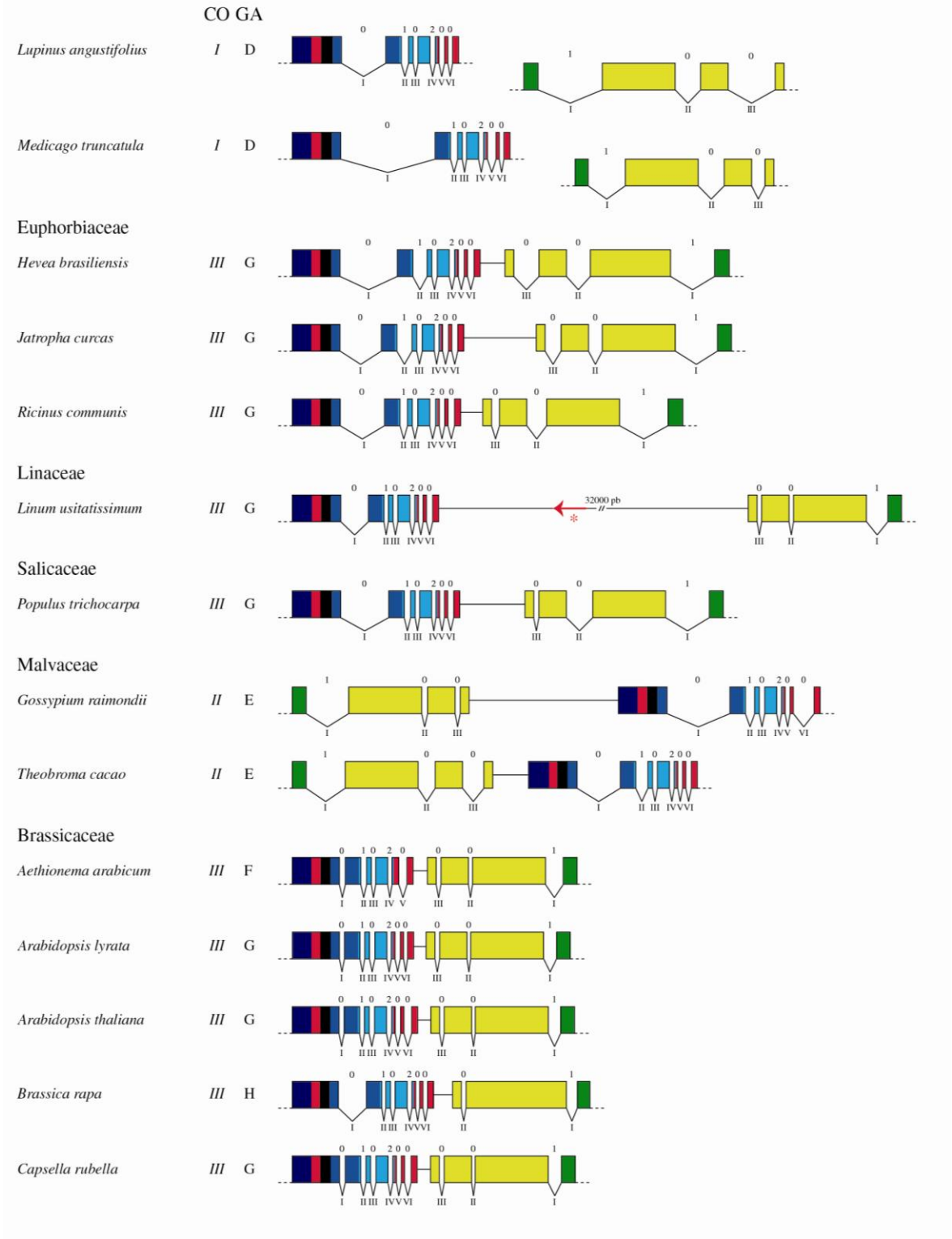
ADDITIONAL FILES

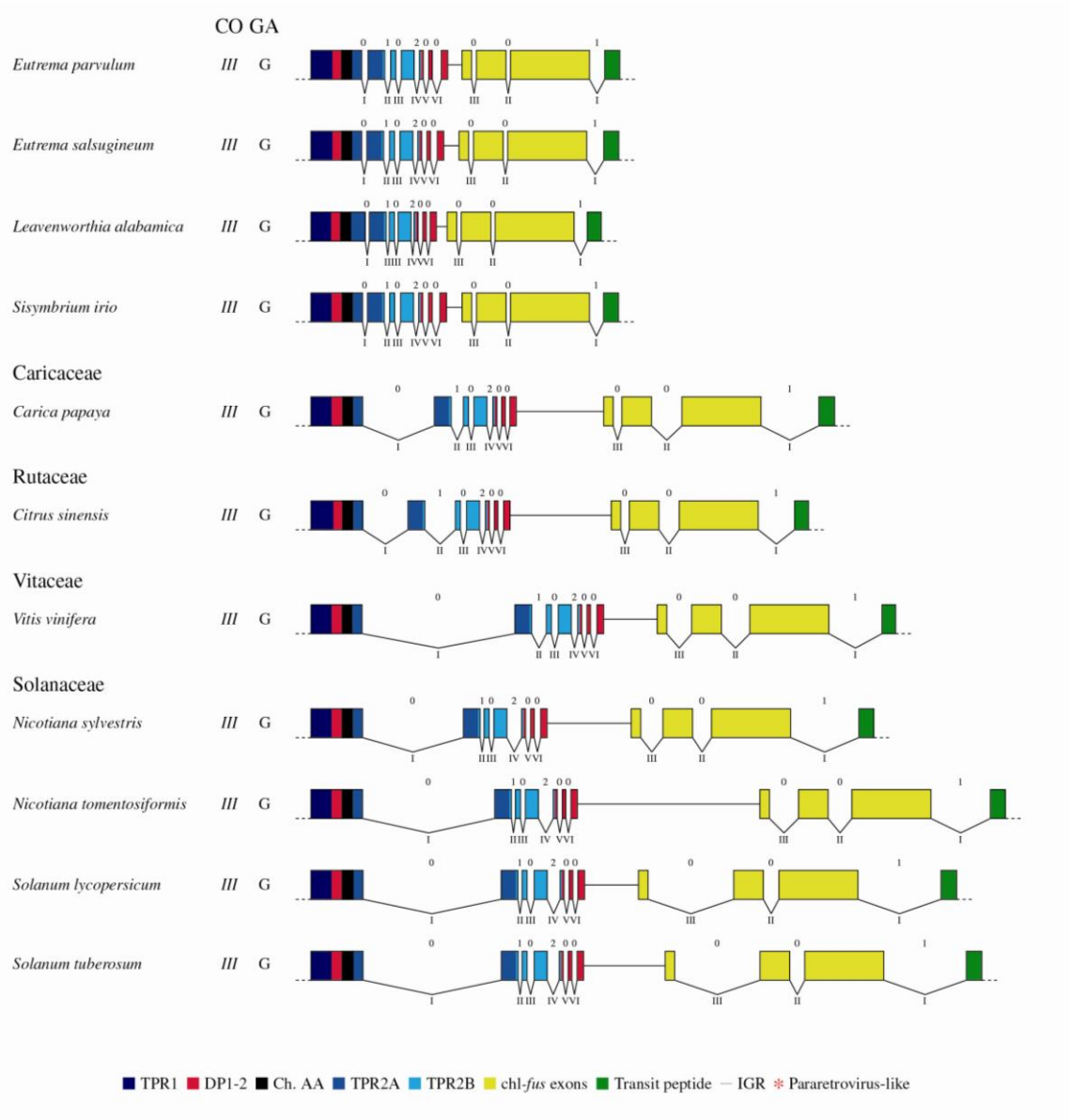
ADDITIONAL FILE A

Figure S1. Detailed gene structure and chromosomal arrangement of the pair of genes *hop* and chl-*fus*, for the 51 plant genomes under study.




CO GA
DICOTS
Cucurbitaceae
Citrullus lanatus
I D
Cucumis melo
I D
Cucumis sativus
I D
Cannabaceae
Cannabis sativa
III G
Moraceae
Morus notabilis
III G
10000 pb
Rosaceae
Fragaria vesca subsp vesca
III G
Malus domestica
III G
Prunus mume
III G
Prunus persica
III G
Pyrus bretschneideri
III G
Fabaceae
Cajanus cajan
III G
Cicer arietinum
I D
Glycine max
III G


CO GA
Lupinus angustifolius I D
Medicago truncatula I D
Euphorbiaceae
Hevea brasiliensis III G
Jatropha curcas III G
Ricinus communis III G
Linaceae
Linum usitatissimum III G
32000 pb
Salicaceae
Populus trichocarpa III G
Malvaceae
Gossypium raimondii II E
Theobroma cacao II E
Brassicaceae
Aethionema arabicum III F
Arabidopsis lyrata III G
Arabidopsis thaliana III G
Brassica rapa III H
Capsella rubella III G



CO: classification by microcolinearity (categories I to III); GA: classification by gene arrangement, according to the exon–intron structure of both combined *hop* and chl-*fus* (categories A to H). Arabic and roman numbers represent intron phase (0, 1, or 2) and succession of introns from I to I+n, respectively; *hop* introns are named as I_h, II_h, III_h, etc., and chl-*fus* introns are named as I_f, II_f, III_f, etc. Exons coding for TPR and DP domains are color-coded according to conventions of Figure 4. Non-syntenic genes are drawn on separate chromosomes.

ADDITIONAL FILE B

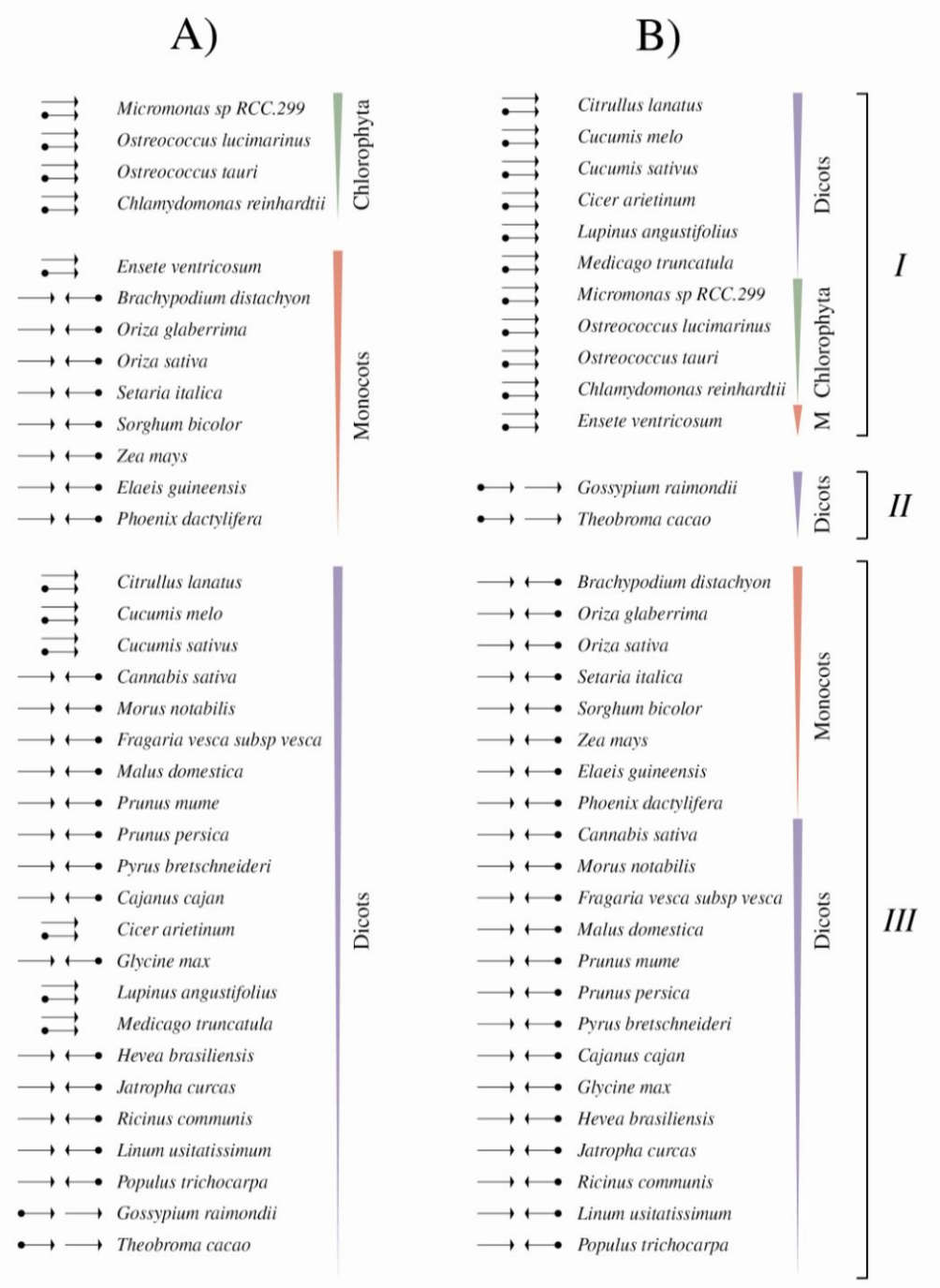
Table S1. Plant species whose *hop* and chl-*fus* genes do not locate on the same chromosome.

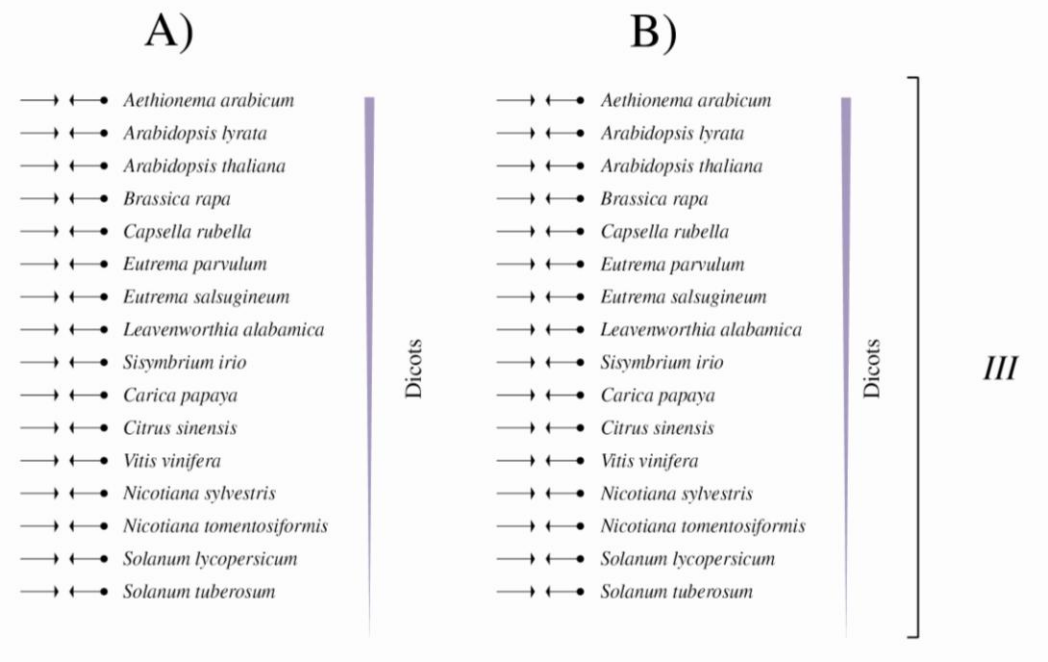
Species	Family	Chl-*fus* Accession Number	Chromosome number	*Hop* Accession Number	Chromosome number
Micromonas sp. RCC299	Mamiellaceae	XP_002500081	2	XP_002500383	3
Ostreococcus lucimarinus	Mamiellaceae	XP_001419031	7	XP_001418158	6
Ostreococcus tauri	Mamiellaceae	XM_003080500	7	XM_003079642	6
Chlamydomonas reinhardtii	Chlamydomonadaceae	XM_001701793	N.A.	XP_001691869	N.A.
Ensete ventricosum	Musaceae	AMZH01015354	N.A.	AMZH01008475	N.A.
Citrullus lanatus	Cucurbitaceae	AGCB01004585	N.A.	AGCB01006484	N.A.
Cucumis melo	Cucurbitaceae	CAJI01003926	N.A.	CAJI01012439	N.A.
Cucumis sativus	Cucurbitaceae	XM_004147564	N.A.	XM_004147890	N.A.
Cicer arietinum	Fabaceae	XM_004515686	8	XM_00451602	N.A.
Lupinus angustifolius	Fabaceae	AOCW01054016	N.A.	AOCW01121688	N.A.
Medicago truncatula	Fabaceae	NC_016410	4	NC_016411	5

N.A., Not Available.

ADDITIONAL FILE C

Figure S2. Graphic representation of microsynteny between *hop* and chl-*fus* genes among all plant species studied.

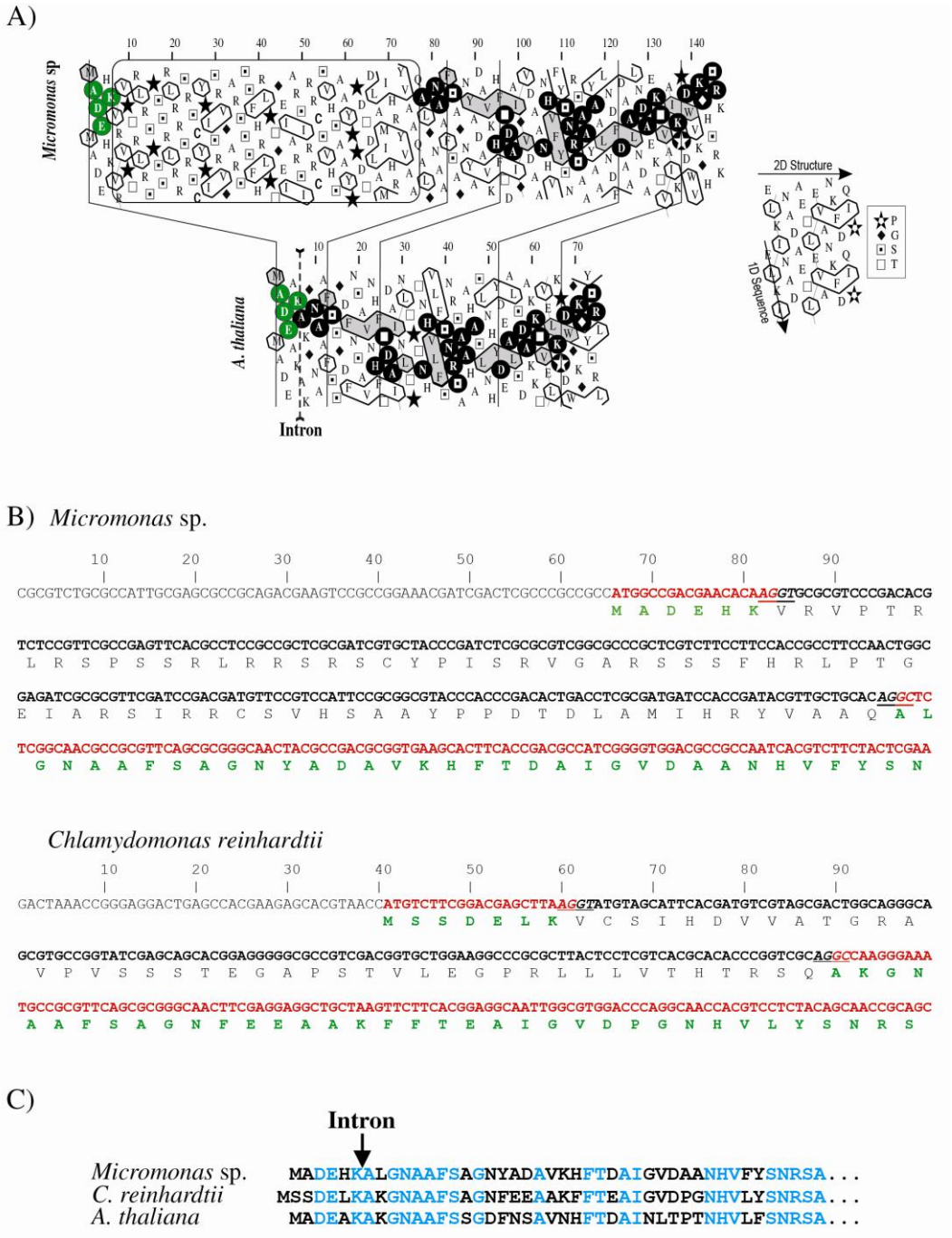




(A) Plant species are ranked in the taxonomic order Chlorophyta, Monocots and Dicots. **(B)** Plant species are ranked by microsyntenic categories I, II and III. Arrows represent the transcriptional orientation of *hop* and chl-*fus* genes.

ADDITIONAL FILE D

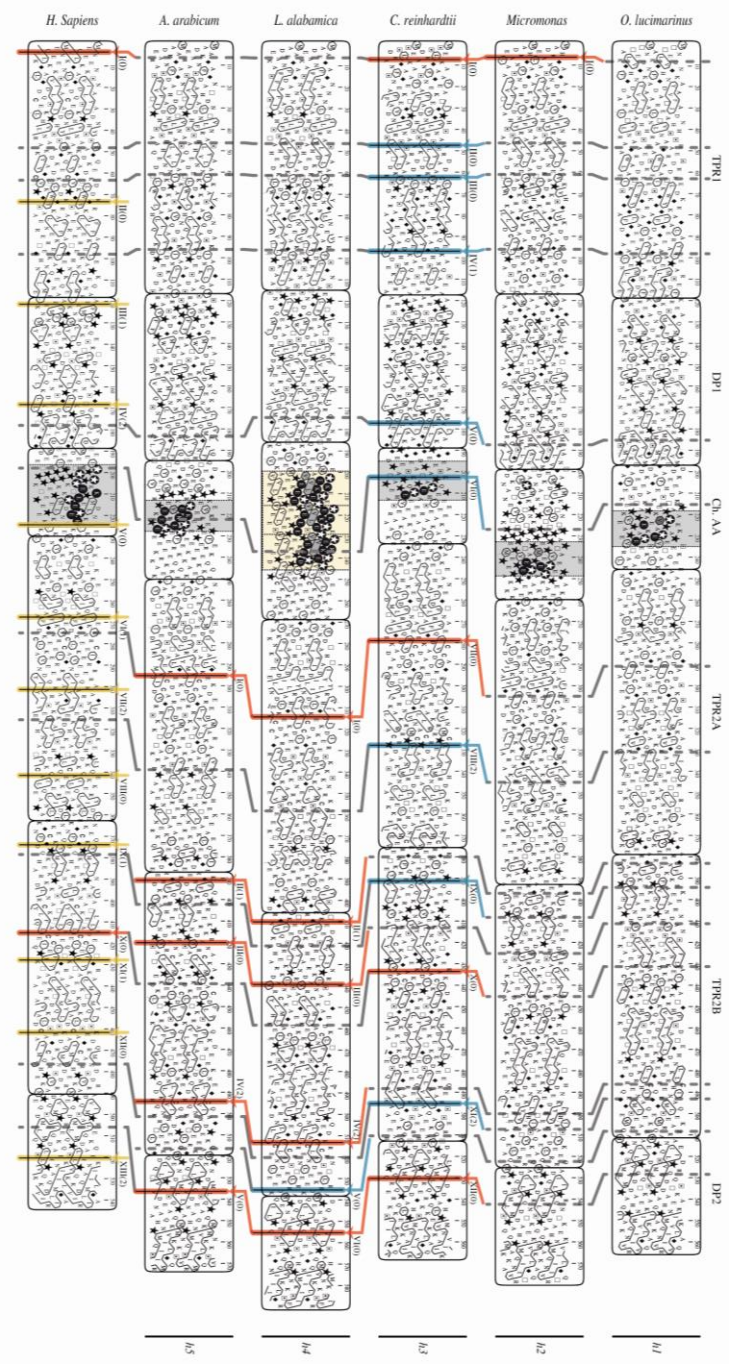
Figure S3. Prediction of an intron (dotted vertical line) in *Micromonas* sp. *hop* gene (GenBank: XP_002500383), downstream of the first seven codons.



(A) HCA alignment of the N-terminal amino acids of *Micromonas* sp. and *A. thaliana* Hop proteins. Extra 71 amino acids in the *Micromonas* sp. Hop protein are bordered by a rounded rectangle. Vertical lines connect analogous positions in both proteins. Conserved hydrophobic clusters are gray shaded. Relevant nonhydrophobic identities are indicated by circles on black background. The way to read the sequence and secondary structures, as well as special symbols, are indicated in the inset. **(B)** Predicted translation of the 5' regions for *Micromonas* sp. and *C. reinhardtii hop* genes. We propose that nucleotides in bold belong to a phase-0 intron, which is in frame with the first and second exons. Splice sites are in italic and underlined. **(C)** ClustalW alignment of the N-terminal amino acids of predicted *Micromonas* sp., *C. reinhardtii* and *A. thaliana* Hop proteins. The arrow indicates the position of the putative intron I_h in *Micromonas* sp., and *C. reinhardtii hop* genes.

ADDITIONAL FILE E

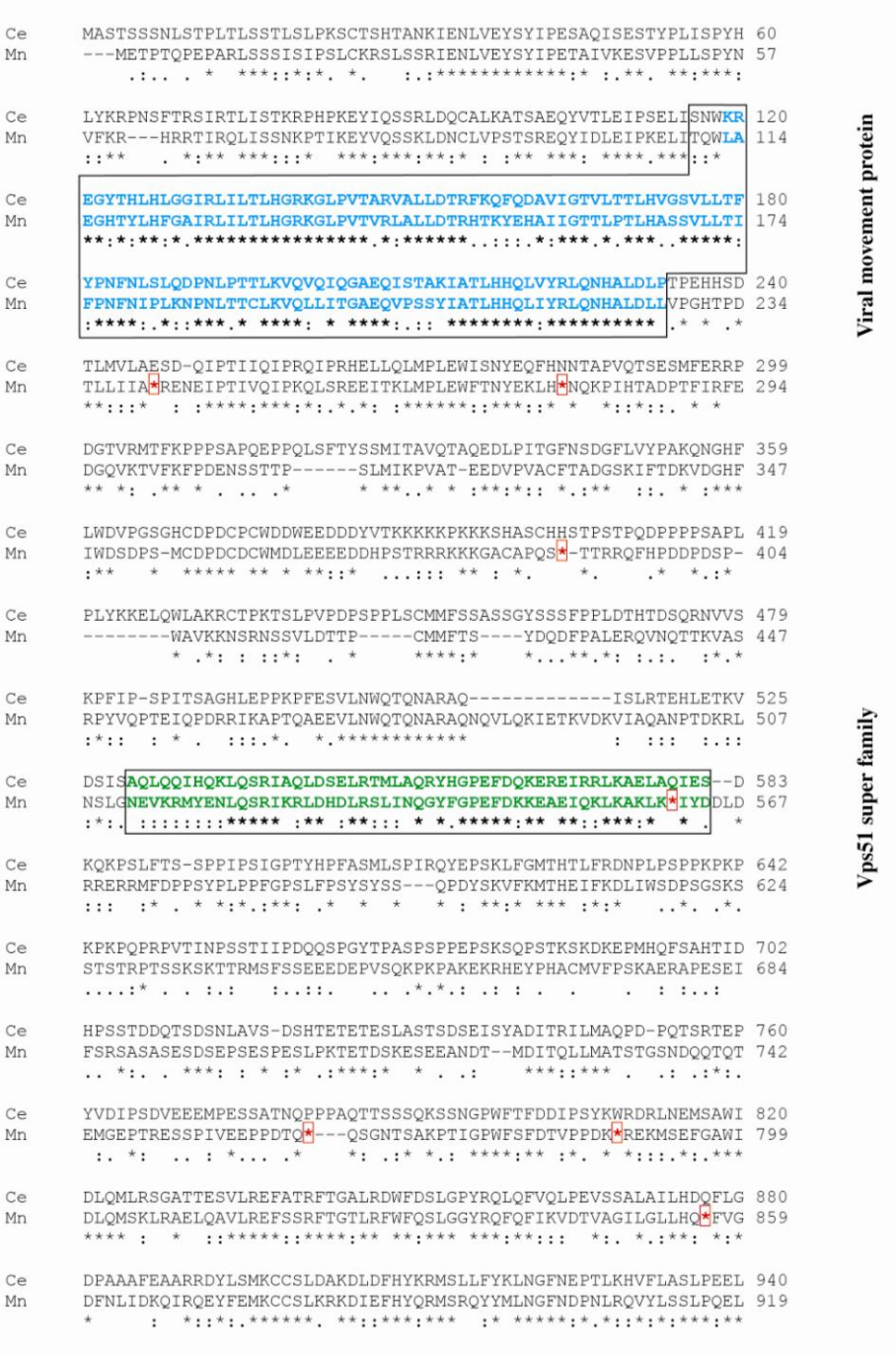
Figure S4. 2D-alignment of plant Hop proteins from members of the five categories of exon–intron organization of *hop* genes (h1 to h5).

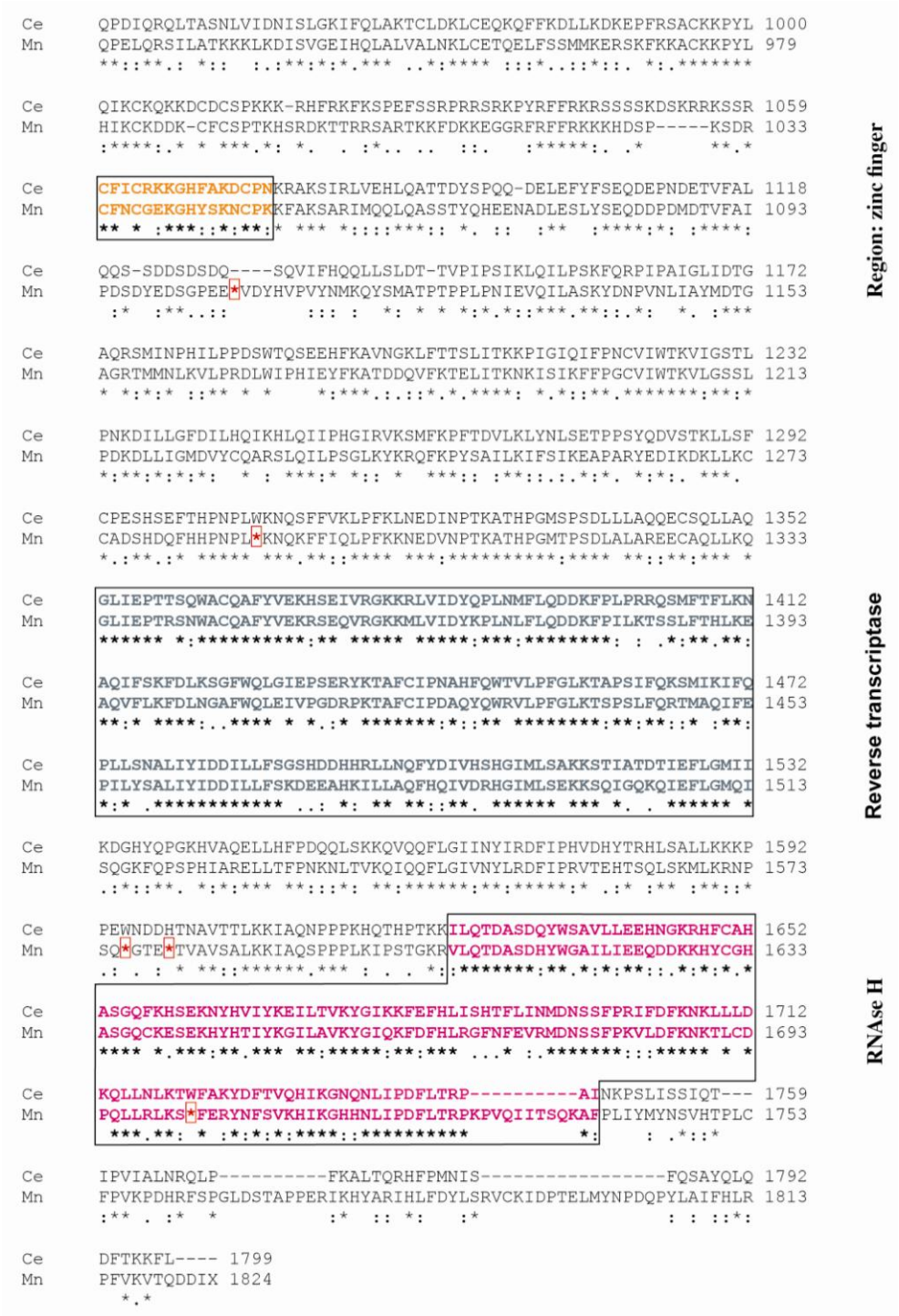


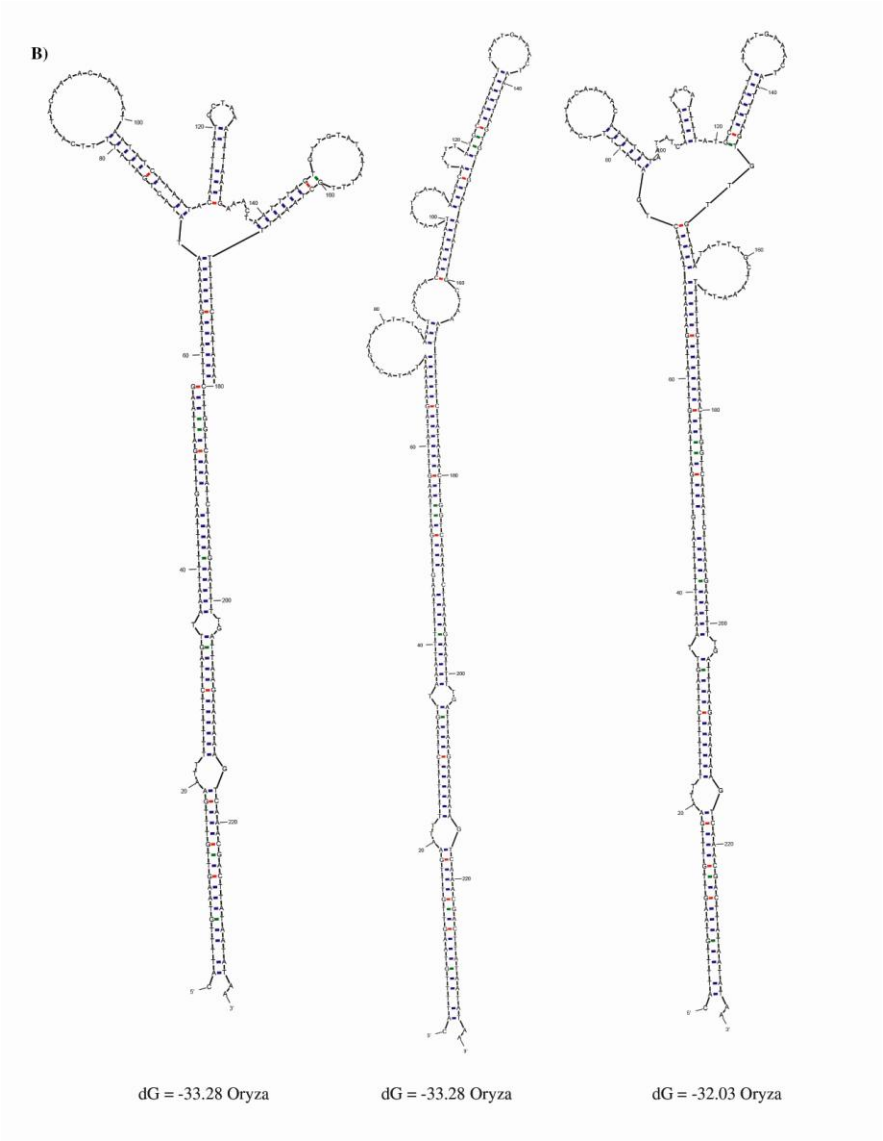
The way to read the sequence and special symbols is the same of Additional file 4: Figure S3 (A). Solid vertical colored lines mark intron positions and dashed lines connect equivalent sites in orthologous proteins. Blue introns: species-specific introns (h3 and h4) introns; red introns: Introns shared among classes h1 to h5; Human Hop protein is represented in the bottom. Yellow introns: Human-specific introns. Gray boxes: strict identities with respect to the *A. alabamica* VPEVEKKLEPEPEP triplet repeat (yellow box). Roman and Arabic numbers represent the succession of introns from I to I+n and intron phase (0, 1, or 2), respectively. TPR, DP and Ch. AA domains are bordered by rectangles with rounded corners. Domain names are on the top.

ADDITIONAL FILE F

Figure S5. Hypothetical genes found within the IGR between the *hop* and chl-*fus* genes.



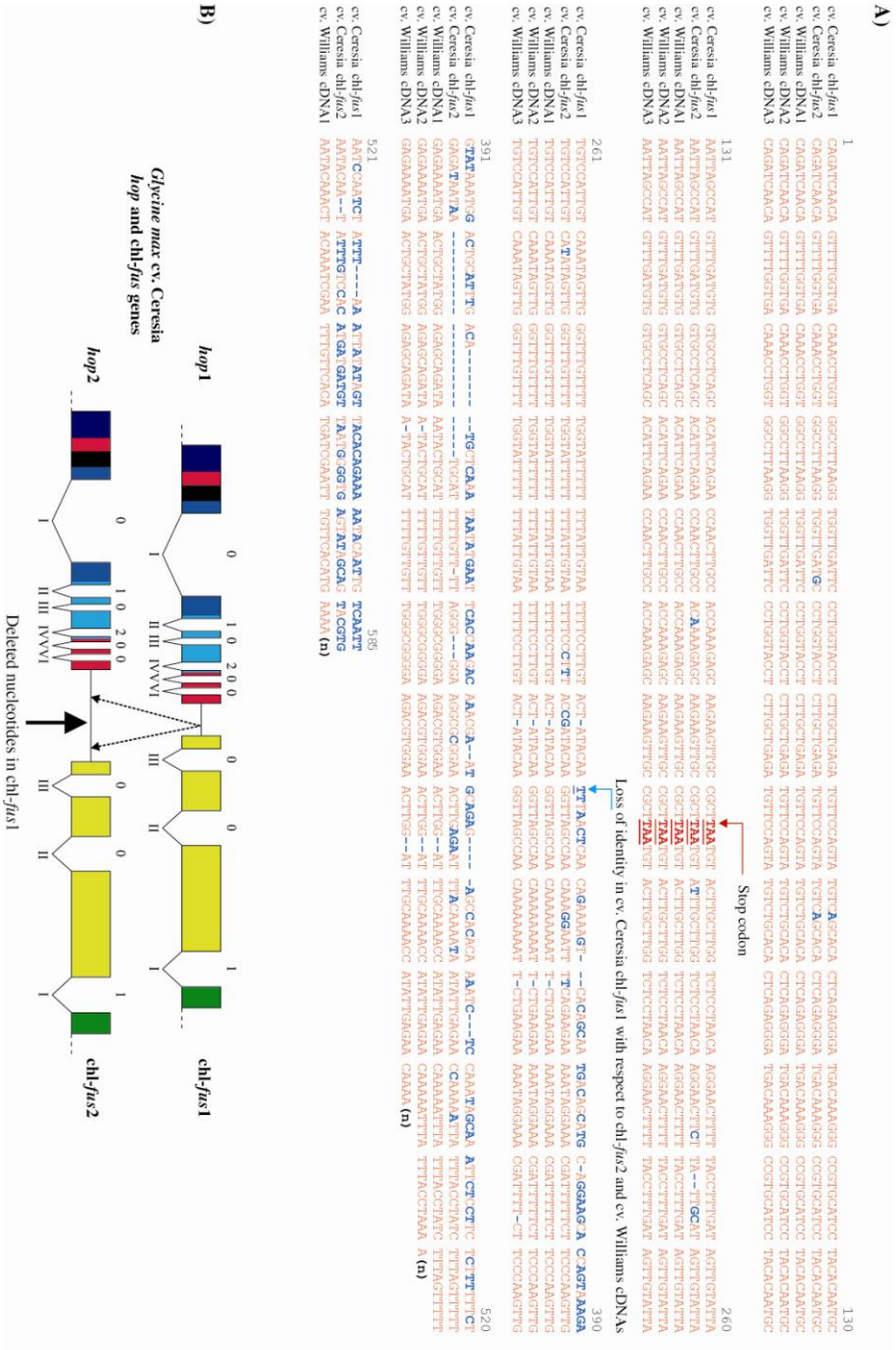

Ce QPDIQRQLTASNLVIDNISLGKIFQLAKTCLDKLCEQKQFFKDLLKDKEPFRSACKKPYL 1000
Mn QPELQRSILATKKKLKDISVGEIHQLALVALNKLCETQELFSSMMKERSKFKKACKKPYL 979
Ce QIKCKQKKDCDCSPKKK-RHFRKFKSPEFSSRPRRSRKPYRFFRKRSSSSKDSKRRKSSR 1059
Mn HIKCKDDK-CFCSPTKHSRDKTTRRSARTKKFDKKEGGRFRFFRKKKHDSP-----KSDR 1033
Ce CFICRKKGHFAKDCPNKRAKSIRLVEHLQATTDYSPQQ-DELEFYFSEQDEPNDETVFAL 1118
Mn CFNCGEKGHYSKNCPKKFAKSARIMQQLQASSTYQHEENADLESLYSEQDDPDMDTVFAI 1093
Ce QQS-SDDSDSDQ----SQVIFHQQLLSLDT-TVPIPSIKLQILPSKFQRPIPAIGLIDTG 1172
Mn PDSDYEDSGPEE*VDYHVPVYNMKQYSMATPTPPLPNIEVQILASKYDNPVNLIAYMDTG 1153
Ce AQRSMINPHILPPDSWTQSEEHFKAVNGKLFTTSLITKKPIGIQIFPNCVIWTKVIGSTL 1232
Mn AGRTMMNLKVLPRDLWIPHIEYFKATDDQVFKTELITKNKISIKFFPGCVIWTKVLGSSL 1213
Ce PNKDILLGFDILHQIKHLQIIPHGIRVKSMFKPFTDVLKLYNLSETPPSYQDVSTKLLSF 1292
Mn PDKDLLIGMDVYCQARSLQILPSGLKYKRQFKPYSAILKIFSIKEAPARYEDIKDKLLKC 1273
Ce CPESHSEFTHPNPLWKNQSFFVKLPFKLNEDINPTKATHPGMSPSDLLLAQQECSQLLAQ 1352
Mn CADSHDQFHHPNPL*KNQKFFIQLPFKKNEDVNPTKATHPGMTPSDLALAREECAQLLKQ 1333
Ce GLIEPTTSQWACQAFYVEKHSEIVRGKKRLVIDYQPLNMFLQDDKFPLPRRQSMFTFLKN 1412
Mn GLIEPTRSNWACQAFYVEKRSEQVRGKKMLVIDYKPLNLFLQDDKFPILKTSSLFTHLKE 1393
Ce AQIFSKFDLKSGFWQLGIEPSERYKTAFCIPNAHFQWTVLPFGLKTAPSIFQKSMIKIFQ 1472
Mn AQVFLKFDLNGAFWQLEIVPGDRPKTAFCIPDAQYQWRVLPFGLKTSPSLFQRTMAQIFE 1453
Ce PLLSNALIYIDDILLFSGSHDDHHRLLNQFYDIVHSHGIMLSAKKSTIATDTIEFLGMII 1532
Mn PILYSALIYIDDILLFSKDEEAHKILLAQFHQIVDRHGIMLSEKKSQIGQKQIEFLGMQI 1513
Ce KDGHYQPGKHVAQELLHFPDQQLSKKQVQQFLGIINYIRDFIPHVDHYTRHLSALLKKKP 1592
Mn SQGKFQPSPHIARELLTFPNKNLTVKQIQQFLGIVNYLRDFIPRVTEHTSQLSKMLKRNP 1573
Ce PEWNDDHTNAVTTLKKIAQNPPPKHQTHPTKKILQTDASDQYWSAVLLEEHNGKRHFCAH 1652
Mn SQ*GTE*TVAVSALKKIAQSPPPLKIPSTGKRVLQTDASDHYWGAILIEEQDDKKHYCGH 1633
Ce ASGQFKHSEKNYHVIYKEILTVKYGIKKFEFHLISHTFLINMDNSSFPRIFDFKNKLLLD 1712
Mn ASGQCKESEKHYHTIYKGILAVKYGIQKFDFHLRGFNFEVRMDNSSFPKVLDFKNKTLCD 1693
Ce KQLLNLKTWFAKYDFTVQHIKGNQNLIPDFLTRP----------AINKPSLISSIQT--- 1759
Mn PQLLRLKS*FERYNFSVKHIKGHHNLIPDFLTRPKPVQIITSQKAFPLIYMYNSVHTPLC 1753
Ce IPVIALNRQLP----------FKALTQRHFPMNIS-----------------FQSAYQLQ 1792
Mn FPVKPDHRFSPGLDSTAPPERIKHYARIHLFDYLSRVCKIDPTELMYNPDQPYLAIFHLR 1813
Ce DFTKKFL---- 1799
Mn PFVKVTQDDIX 1824
Region: zinc finger
Reverse transcriptase
RNAse H



(A) ClustalW alignment between an inferred pseudogen encoded by the *M. notabilis* IGR (Mn) and a *Citrus endogenous* (Ce) pararetrovirus (Genbank:KF800044). Colored boxes represent signature domains, including the viral movement protein, zinc finger, reverse transcriptase, and RNAse H. (*): internal stop codons. **(B)** Three predicted secondary structures [73, 74] of the inverted repeat sequences of a Miniature Inverted–Repeat Transposable Element (MITE) found within the IGR of *Oryza* spp.

ADDITIONAL FILE G

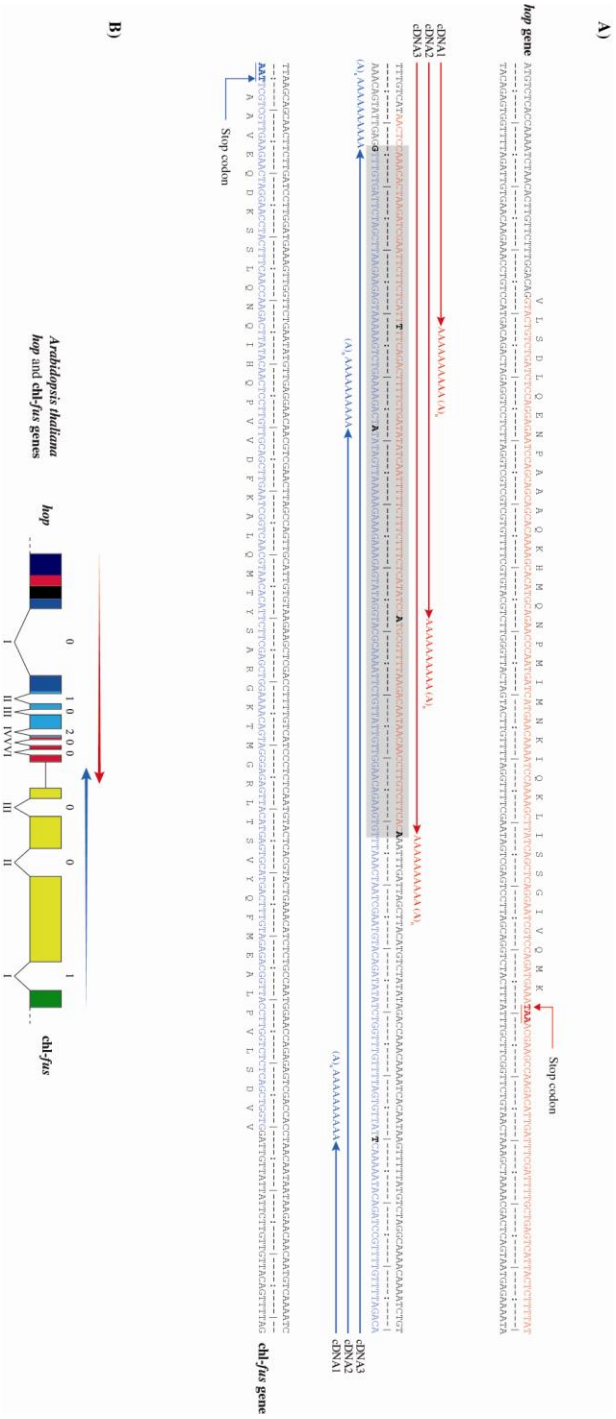
Figure S6. The IGR between the *hop*1 and chl-*fus*1 genes of *G. max* cv. Ceresia is shorter than that of *hop*2 and chl-*fus*2.



(A) Multalin multiple alignment of the 3' region of *G. max* cv. Ceresia chl-*fus*1 and chl-*fus*2 genes with three *G. max* cv. Williams cDNAs. Translational termination stop codons (TAA) are bold and underlined (red arrow). Blue nucleotides in chl-*fus*1 and chl-*fus*2 genes: Mismatched positions with respect to cDNAs. Identity between chl-*fus*1 and chl-*fus*2 + cDNA sequences stop 123 positions downstream of the stop codon (blue arrow). A(n): poly-A tail. **(B)** Structure of the two genetic loci consisting each of a pair of *hop* and chl-*fus* genes, in *G. max* cv. Ceresia. Note that *hop* and chl-*fus* genes keep opposite polarity. Vertical arrows indicate deleted nucleotides (ca. 680 bp) in chl-*fus*1. Intron number and phase are the same of Fig. 4.

ADDITIONAL FILE H

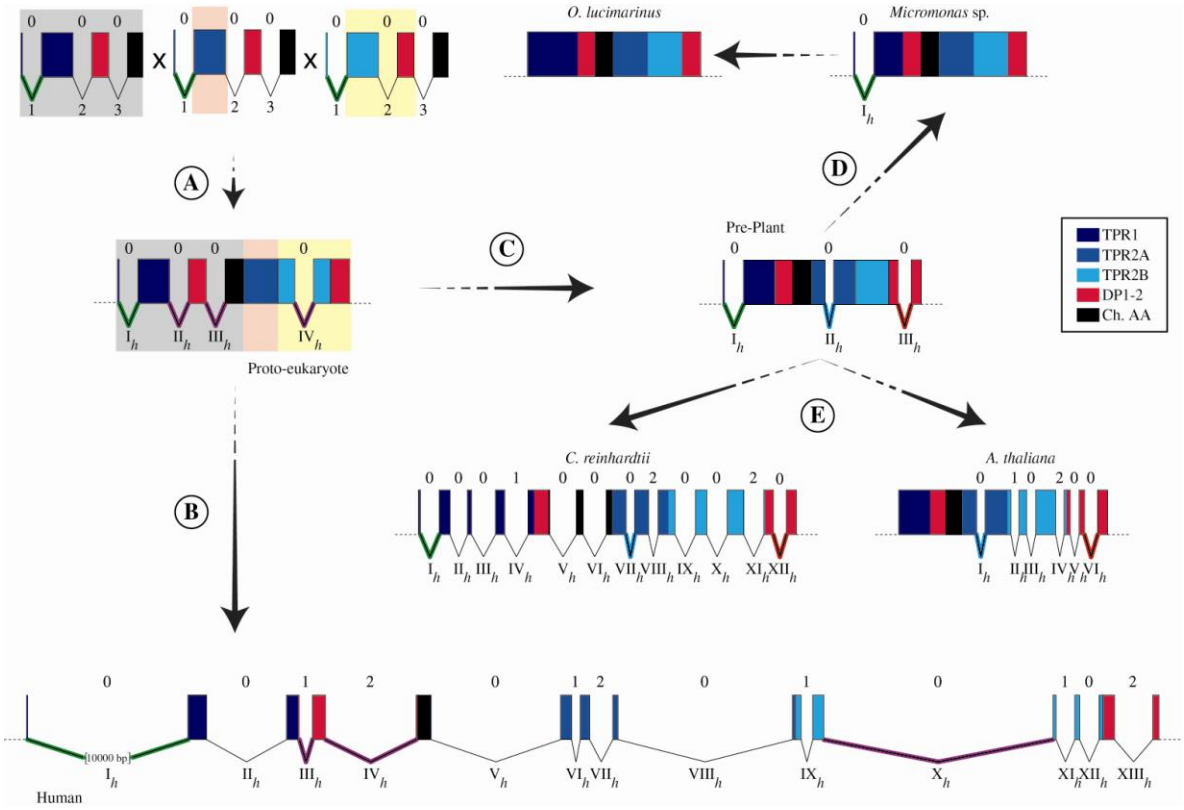
Figure S7. In *A. thaliana*, the *hop* and chl-*fus* genes become overlap in the 3' end.



(A) Graphic view of the IGR separating the *hop* and chl-*fus* genes in *A. thaliana*. Last exons and 3' non-coding ends are color-coded: red, *hop* gene; blue, chl-*fus* gene. The long horizontal arrows represent retrieved cDNAs from Genbank (see Methods for accession numbers). The shaded box covers the overlapping 3' non-coding cDNA ends. (A)n: poly-A tails. **(B)** Topology of the *hop* and chl-*fus* genes, showing the absence of IGR region and overlapping 3' ends.

ADDITIONAL FILE I

Figure S8. Hypothetical evolutionary model of the *hop* gene.



(A) Inside the nucleus of the primitive eukaryote, successive recombinations of a primary «mini-exon − phase-0 intron – TPR domain − phase-0 intron − DP domain − phase-0 intron – Ch. AA » module led to the formation of a 'proto-eukaryote *hop* gene'. Blue, pink and yellow boxes enclose remaining exons and introns. Through the modular assembly of the 'proto-eukaryote *hop*', two phase-0 introns remained (one green, one purple) **(B)** Evolution from the 'proto-eukaryote form' to the present-day human *hop* gene. The green and purple phase-0 introns were preserved. Furthermore, eleven new introns were gained in the process. **(C)** The 'proto-eukaryote form' evolved to 'pre-plant form'. The purple intron was lost, leading to the fusion of the DP1 and TPR2A domains; meanwhile, the blue and red introns were gained. **(D)** The 'pre-plant form' gradually reduced its intron number to

zero, giving rise to contemporary *Micromonas* sp. and *O. lucimarinus hop* genes. **(E)** Nevertheless, on the way to the evolution towards more complex photosynthetic eukaryotes, the ‘pre-plant form’ eventually acquired a broad number of new introns such as in *C. reinhardtii* or higher plants (e.g., *A. thaliana*), but conserving the blue and red introns.