Euglena Light-Harvesting Complexes Are Encoded by Multifarious Polyprotein mRNAs that Evolve in Concert

Adam G. Koziol and Dion G. Durnford

Department of Biology, University of New Brunswick, Fredericton, New Brunswick, Canada

Light-harvesting complexes (LHCs) are a superfamily of chlorophyll- and carotenoid-binding proteins that are responsible for the capture of light energy and its transfer to the photosynthetic reaction centers. Unlike those of most eukaryotes, the LHCs of *Euglena gracilis* are translated from large mRNAs, producing polyprotein precursors consisting of multiple concatenated LHC subunits that are separated by conserved decapeptide linkers. These precursors are posttranslationally targeted to the chloroplast and cleaved into individual proteins. We analyzed expressed sequence tags from *Euglena* to further characterize the structural features of the LHC polyprotein-coding genes and to examine the evolution of this multigene family. Of the 19 different LHC transcriptional units we detected, 17 encoded polyproteins composed of both tandem and nontandem repeats of LHC subunits; organizations that likely occurred through unequal crossing-over. Of the 2 nonpolyprotein-encoding LHC transcripts detected, 1 evolved from the truncation of a polyprotein-coding gene. Duplication of LHC polyprotein-coding genes was particularly important in the LHCI gene family where multiple paralogous sequences were detected. Intriguingly, several of the individual LHC-coding subunits both within and between transcriptional units appeared to be evolving in concert, suggesting that gene conversion has been a significant mechanism for LHC evolution in *Euglena*.

Introduction

The production of polyproteins is commonly associated with viruses where the viral genome is transcribed into a single mRNA molecule and translated into a large polyprotein. This polyprotein is then cleaved into individual proteins with distinct functions, often by a protease contained within the polyprotein itself (Carrington and Dougherty 1987). Eukaryotic polyproteins, however, are uncommon, and they often possess a number of tandemly repeated protein subunits that have identical or related functions rather than proteins of unrelated functions as with viral polyproteins. The polyubiquitin gene, for instance, is commonly observed in eukaryotes, and this polyprotein contains a variable number of tandemly repeating, nearly identical ubiquitin units that are processed into individual proteins (Ozkaynak et al. 1984; Arribas et al. 1986; Graham et al. 1989). Profilaggrin, involved in the aggregation of keratin intermediate filaments during mammalian epidermal differentiation, is also expressed as a polyprotein precursor consisting of tandemly repeating units (Gan et al. 1990). Other examples include the lipid-binding proteins of nematodes that contain up to 10 repeating units (Kennedy 2000) and the antifreeze glycoproteins of Antarctic fish (Chen et al. 1997). In most cases, with the exception of polyubiquitin, the individual units of the polyprotein precursors are separated by polylinkers that act as processing sites (Rowan et al. 1996; Chen et al. 1997).

Euglena and many dinoflagellates are unique in that they contain chloroplasts that are surrounded by 3 membranes, though they obtained their plastids through independent secondary symbiotic events (Cavalier-Smith 1999). Additionally, both *Euglena* and dinoflagellates encode multiple nuclear-encoded plastid-targeted proteins translated as polyprotein precursors. These polyproteins are routed through the endomembrane system (Osafune et al. 1990) and directed to the chloroplasts via complex N-terminal targeting sequences (Durnford and Gray 2006). Once in the chloroplast, the individual proteins are liberated through proteolytic cleavage of conserved decapeptide linkers (polylinkers) (Muchhal and Schwartzbach 1992; Hiller et al. 1995).

In this study, we examined the complexity of the *Euglena* light-harvesting complex (LHC) gene family, which is known to produce LHCs as large polyproteins that are subsequently processed into individual units with the chloroplast (Houlné and Schantz 1987, 1988; Muchhal and Schwartzbach 1992). These polyproteins have previously been classified as LHCI or LHCII based upon whether their LHC subunits are predicted to associate exclusively with photosystem I (Houlné and Schantz 1988) or with photosystem II (Houlné and Schantz 1987; Muchhal and Schwartzbach 1992). Recently, we examined the antenna complexity and evolution in several Chl *a*/*b*-containing organisms and discovered a very diverse LHCII family in *Euglena* (Koziol et al. 2007). Here, we specifically examine the complexity of LHC polyproteins to determine LHC-subunit diversity, coding unit structure, and evolution of *Euglena* LHC units within individual polyprotein-encoding cDNAs.

Materials and Methods

cDNA Libraries and Data Mining

cDNA libraries from *Euglena gracilis* (strain Z) were commercially prepared from RNA isolated under a variety of growth conditions. Bacterial plating, picking, DNA preparation, sequencing, trace processing, and data mining have been previously described in Koziol et al. (2007). Individual expressed sequence tags have been deposited in dbEST National Center for Biotechnology Information and the annotated clusters have been deposited in GenBank (table 1). The LHC sequence data were obtained from clustering 25,595 individual EST reads, and because many of the clusters contained related or nearly identical coding units, we confirmed the clustering in a number of ways. This included confirming the sizes of individual ESTs within a cluster by restriction mapping the largest cDNA and manually sequencing individual cDNAs in areas of limited sequence coverage.

Key words: light-harvesting complex, concerted evolution, polyprotein, *Euglena*.

E-mail: durnford@unb.ca.

Mol. Biol. Evol. 25(1):92–100. 2008
doi:10.1093/molbev/msm232
Advance Access publication October 18, 2007

© The Author 2007. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

Table 1
***Euglena* LHC Transcript Nomenclature and Summary Data**

Gene Name	Cluster ID	GenBank Accession Number	EST Frequency	Total Clustered Sequence (kb)	Transcript Size (kb) (Northern Blot)	Number of Protein Subunits (Estimate)
Lhca1	EEL4043	BK005985	45	3.74	3.9	5
Lhca2	EEL3996	BK005984	27	3.69		5
Lhca3	EEL3911	BK005981	15	1.61		1
Lhca4	EEL3716	EU124874	7	2.90		(5)
Lhca5	EEL4008	EU124875	30	3.33		5
Lhca6	EEL3628	EU124876	6	1.52		(5)
Lhca7	EEL1079	EU124877	1	1.03		(5)
Lhca8	EEL2369	EU124878	2	2.12		(5)
Lhca9	EEL3454	EU124879	4	1.71		?
Lhca10	EEL3771	EU124880	8	2.76	4.2	(6)
Lhca11	EEL9503	EU124881	4	1.62		?
Lhcbm1	EEL3904	BK005978	14	2.70	6.6[a]	(8)
Lhcbm3	EEL1433	BK005979	1	1.01		?
Lhcbm4	EEL3244	BK005980	3	1.28		?
Lhcbm5	EEL3944	BK005982	17	2.39	7.5[b]	(9)
Lhcbm6	EEL3893	BK005983	13	2.57	4.9	(6)
Lhcbm7	EEL9478	EU124884	6	1.76		?
Lhcbm8	EEL3814	EU124882	10	2.09	5.7	(7)
Lhcbm9	EEL3827	EU124883	8	1.84		?
Lhcbm10	EEL2468	EU124885	2	2.10	3.5	(4)
Lhcb4	EEL4056	BK005977	101	1.44	1.4	1

NOTE.—?,undetermined.
[a] As determined in Muchhal and Schwartzbach (1992).
[b] As determined in Houlné and Schantz (1987).

Phylogenetic Analyses

Phylogenetic analyses of LHC-nucleotide sequences were performed using Bayesian and maximum likelihood tests. MrModeltest 2.2 (Nylander 2004) was used to select the best-fit model of nucleotide substitution for each of the analyses for the different alignments based upon the Akaike Information Criterion framework (Akaike 1974). The PHYML program (Guindon and Gascuel 2003; Guindon et al. 2005) was used for the maximum likelihood analyses (http://atgc.lirmm.fr/phyml/) utilizing the general time reversible (GTR) nucleotide substitution matrix, with gamma correction (6 categories), and accounting for the number of invariant sites.

MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001) was used for the Bayesian analyses. GTR with 4 gamma distribution categories and incorporating the number of invariant sites was used as the substitution matrix. The data set was partitioned by codon to allow for different rates of change in each codon position. The number of generations performed was 5.00×10^6, with a sampling frequency of 100 and a 25% burn-in value. The consensus type was allcompat and posterior probabilities that support a node on the resulting consensus tree (>0.50) are shown. Bayesian inference was conducted using the resources of the Computational Biology Service Unit from Cornell University (http://cbsuapps.tc.cornell.edu/mrbayes.aspx). All phylogenetic trees are displayed using TreeView (Page 1996). Untrimmed alignments are available in the Supplementary Material online.

Nucleotide Substitution Analysis

The number of synonymous and nonsynonymous substitutions per site for several of the polyprotein-coding sequences was calculated using the K-Estimator program (Comeron 1999) with the Kimura 2-parameter model correcting for multiple hits and the confidence intervals (CIs) were calculated for the analyses. We also used GENECONV v1.81 (Sawyer 1999) to test for gene conversion events using statistical tests (Sawyer 1989). A mismatch penalty of 1 was used, and 10,000 random permutations of the polymorphic sites were executed in order to evaluate the significance of putative gene conversions. Sequence logo displays of the polylinkers were generated using the online program WebLogo (weblogo.Berkeley.edu/logo.cgi) (Crooks et al. 2004).

Northern Blot Analyses

To determine transcript size, individual cDNAs that constituted specific clusters were selected as probes for Northern hybridization such that the entire 3′ untranslated regions (UTRs) plus a portion of the coding sequence preceding it were included. Single cDNA clones for select LHC-coding genes were isolated, and inserts were excised from the pCDNA3.1(+) vector by an *EcoRI*/*XhoI* double digest and purified from the gel (QIAquick gel extraction kit, Qiagen Mississauga Ontario, Canada). Probes were generated for *Lhcbm8* (2.1 kb fragment from cDNA ELE00009515), *Lhca1* (1.6 kb fragment from cDNA ELE00007187), and *Lhcb4* (1.5 kb fragment from cDNA ELE00008952). Probes were labeled using dCTP 5′-[α-^{32}P], hybridized to the blots, and detected as previously described (Durnford et al. 2003).

Results

Polyprotein Organization and Evolution

We identified 3 polyprotein categories for both LHCII and LHCI: 1) complex proteins of divergent subunits, 2)

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

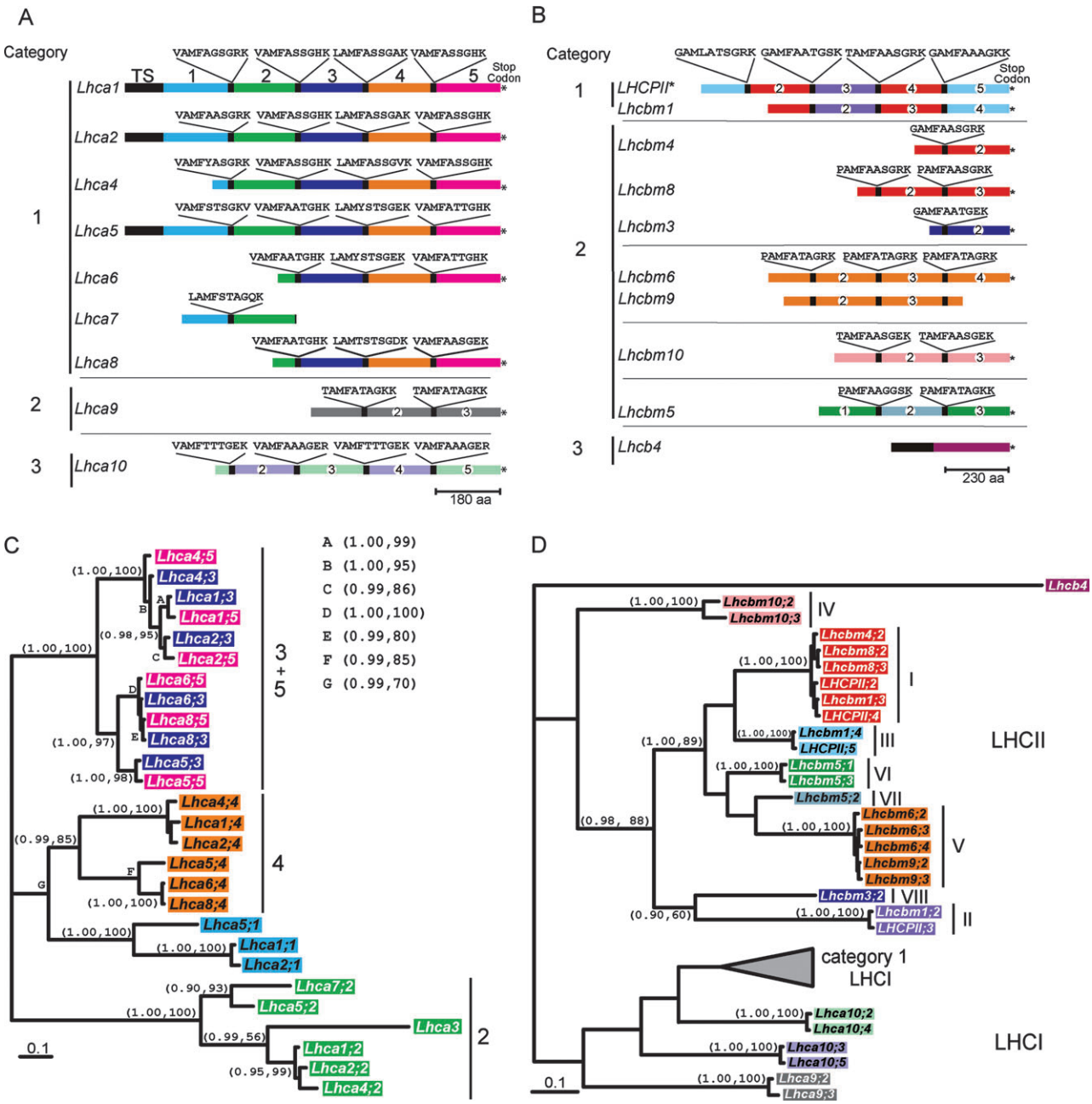

FIG. 1.—*Euglena* polyprotein organization. (*A*) Subunits of LHCI polyprotein-coding genes. Subunits are represented by different colors and usually indicate phylogenetic types. Sequences of the decapeptide linkers are shown for each polyprotein. (*B*) Subunits of LHCII polyprotein-coding genes. Different colors indicate phylogenetic types. Decapeptide linkers are shown for each polyprotein, except for *Lhcbm1* and *Lhcbm9*, which are identical to the sequences that precede them. The LHCPII sequence represents the previously described *Euglena* LHCII (GenBank accession number: X61361). For both figures (*A*) and (*B*), numbers within the LHC subunits indicate the position of the subunit within the polyprotein and are represented by the number following the ";" in the sequences in the phylogenetic analysis. (*C*) Phylogenetic reconstruction of the *Euglena Lhca* gene complement. Colors are the same for the subunits in (*A*). A MrBayes tree is shown ($-\ln L = -5{,}390.502$, $\alpha = 1.158$) with the support values for the MrBayes and maximum likelihood analyses shown at specific nodes. A total of 507 characters were included, and the proportion of invariable sites was 0.139. The average standard deviation (SD) of the split frequencies was 0.001484. A total of 27 sequences were included in the analyses. Numbers refer to the subunit in the polyprotein as labeled in (*A*). (*D*) Phylogenetic reconstruction of the *Euglena LHC* gene complement. Colors are the same for the subunits as in (*B*). Roman numerals denote proposed LHC protein "types." A MrBayes tree is shown ($-\ln L = -7{,}604.465$, $\alpha = 1.220$) with the support values as indicated in (*C*). A total of 378 characters were included, and the proportion of invariable sites was 0.089. The average SD of the split frequencies was 0.005010. A total of 59 sequences were included in the analyses, and the accession numbers are indicated in table 1.

tandem duplications of nearly identical subunits, and 3) tandem duplication of pairs of divergent subunits (figs. 1*A* and *B*). The previously described *Euglena* LHCI and LHCII polyproteins (Houlné and Schantz 1988; Muchhal and Schwartzbach 1992) fall within the first category. We found evidence for extensive gene duplication for the category 1 LHCI polyproteins with 7 paralogs (*Lhca1*, *2*, *4–8*; fig. 1*A*). The full-length LHCI category 1 polyproteins contained 5

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

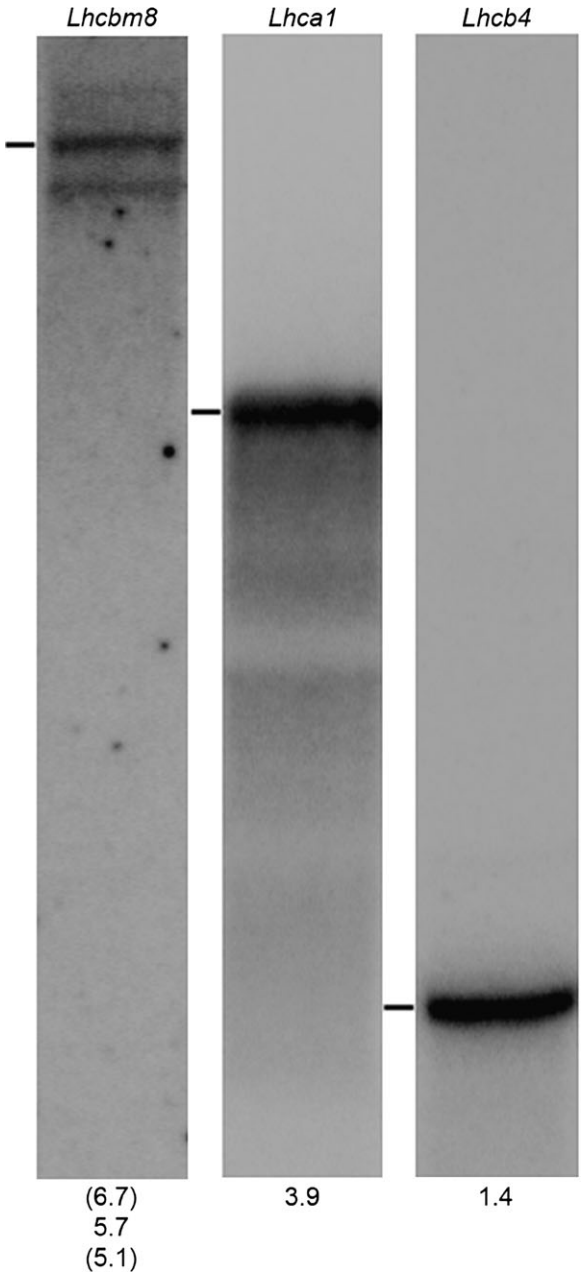

FIG. 2.—Northern hybridization using 20 μg of *Euglena* total RNA with *Lhcbm8*, *Lhca1*, and *Lhcb4* probes. The sizes (Kb) of prominent and minor (sizes in parentheses) hybridizing bands are indicated under each blot.

LHC-coding units per mRNA (fig. 1*A*), and the 3.9-kb transcript size was confirmed by Northern hybridization (fig. 2, *Lhca1*). Phylogenetic analysis of the LHCI category 1 polyproteins, however, indicated that the individual subunits encode divergent members of the LHC gene family with 4 of the 5 subunits falling into distinct subunit "types" (fig. 1*C*; 1, 2, 4, 3/5).

We observed only a single LHCII category 1 polyprotein. This polyprotein was identical to the partial sequence found in GenBank (X61361; Muchhal and Schwartzbach 1992), with the exception of a few mismatches that may be strain specific. The LHCII polyproteins are generally much larger than the LHCI polyproteins, and previous estimates have indicated that the transcript for the category 1 LHCII is 6.6 kb (Muchhal and Schwartzbach 1992). Taking into consideration the average size of the targeting sequence, the UTRs, and the size of the individual subunits, this would give a polyprotein with approximately 8 subunits (table 1). However, due to the large size of the transcripts, all the LHCII cDNAs in our libraries were truncated, and we were only able to detect a total of 4 subunits coded by *Lhcbm1*. For the single LHCII category 1 polyprotein, at least 3 distinct types were well supported (fig. 1*D*, I–III), a number that is likely to increase once the full-length sequence is available.

There was a single, divergent category 2 polyprotein-coding transcript with a weak association to the LHCI clade that had at least 3 tandemly repeated subunits (*Lhca9*) (fig. 1*A* and *D*). In comparison, we found numerous LHCII category 2 polyproteins, including *Lhcbm3*, *4*, *6*, *8–10* (fig. 1*B*). The subunits encoded by these transcripts also cluster within the main LHCII clade, excluding *Lhcbm10* (fig. 1*D*). The estimates of the number of subunits per polyprotein are variable, ranging from 6 to 8 depending on the transcript (table 1). The transcript size for *Lhcbm8* is 5.7 kb (fig. 2), which corresponds to approximately 7 subunits. Interestingly, the tandemly repeating subunits in transcripts *Lhcbm4* and *Lhcbm8* are nearly identical to 2 of the subunits in the category 1 polyprotein-coding gene *LHCPII/Lhcbm1*, and these form a supported clade on the LHCII tree (fig. 1*D*, type I). This close relationship is also apparent by weak hybridization to a 6.7-kb hybridization band detected with the *Lhcbm8* probe (fig. 2), which corresponds to the main *Lhcbm1* transcript size (Muchhal and Schwartzbach 1992). The 5.1-kb band detected with the *Lhcbm8* probe (fig. 2) is likely from the *Lhcbm4* transcript.

Lhca10 provides an example of a tandem arrangement of pairs of LHC subunits defining category 3 polyproteins. *Lhca10* has at least 2 repeats of a pair of LHCI proteins (fig. 1*A* and *D*) and a transcript size of 4.2 kb (table 1), thus containing an estimated 6 subunits. The LHCII-related *Lhcbm5* gene was also labeled a category 3 polyprotein by analogy to *Lhca10*, though we have insufficient sequence data to confirm the organization.

The unexpectedly high nucleotide sequence identity between the coding units of different polyprotein transcripts and between subunits within a polyprotein suggests that they are evolving in concert. For the LHCI category 1 sequences, it is clear that subunits 1, 2, and 4 from different paralogs form supported clades (fig. 1*C*). However, subunits 3 and 5 within each polyprotein gene are more similar to each other than they are to the paralogous subunits from different genes (fig. 1*C*, 3 + 5). We tested the hypothesis that these subunits were evolving in concert by calculating the synonymous (*Ka*) and nonsynonymous (*Ks*) changes per site between the sets of similar LHC subunits. Between the paralogous, yet divergent members of the LHCI-coding family, *Lhca1* and *Lhca5*, there were significantly lower nonsynonymous and synonymous changes per site between subunits 3 and 5 within the polyproteins than compared with the paralogous subunits (3 vs. 3; 5 vs. 5) between the polyproteins (fig. 3, CI = 1%). Though the paralogous subunits between polyprotein genes continue to evolve

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

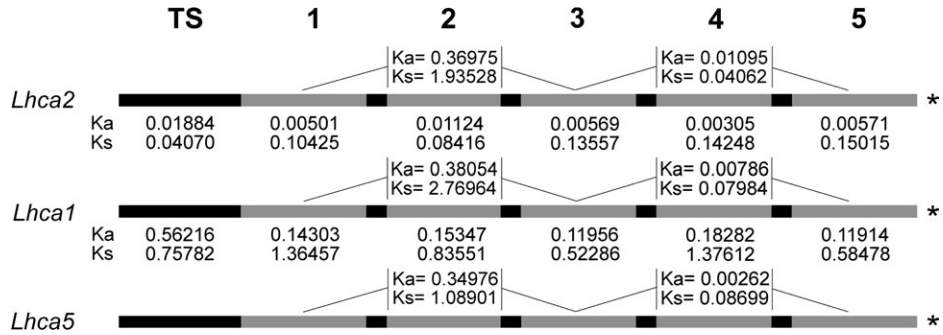

FIG. 3.—Comparison of synonymous (*Ka*) and nonsynonymous (*Ks*) nucleotide substitutions between paralogous subunits of *Lhca1* and *Lhca2* plus *Lhca1* and *Lhca5*. *Ka* and *Ks* values were also determined between subunits 1 and 3 as well as 3 and 5 for each polyprotein and are indicated by connecting lines. The *Ka* and *Ks* values for the plastid targeting sequence (TS) are also shown.

independently of one another, subunits 3 and 5 within each paralogous gene appear to be evolving in concert. The remaining subunits (1, 2, and 4) had no significant intrapolyprotein subunit similarity and appeared to evolve independently of other subunits in the polyprotein. Moreover, homogenization events between subunits 3 and 5 within the LHCI polyproteins were predicted by GENECONV ($P < 0.02$) (Sawyer 1999). There is also evidence for concerted evolution of LHCII subunits, but in this case, the phenomenon extends to subunits of different transcripts in addition to those within the same transcript. Specifically, subunits 2 and 4 of *Lhcbm1* (LHCPII) and the tandemly arranged subunits of *Lhcbm4* and *8* together form a strongly supported type I clade in the LHCII tree (fig. 1*D*, I). All LHCII type I subunits had low *Ka* and *Ks* values (less than 0.00762 and 0.07128, respectively). Furthermore, using GENECONV, we found that there was strong evidence for these subunits evolving in concert ($P < 0.05$).

The decapeptide linkers separating the coding units within the polyproteins all share similar features that can be divided into 3 sections: a front portion consisting of a variable (small, hydrophobic) residue; an invariant Ala, Met, and an aromatic amino acid (Phe or Tyr); a central portion consisting of 3 small and/or hydroxylated amino acids (Ala, Ser or Thr); and a terminal portion that consists of a conserved Gly, followed by a charged amino acid, and a basic residue (Lys or Arg) (fig. 4). There is high sequence conservation between linkers that precede identical or nearly identical subunits. For instance, all polylinkers in *Lhcbm6* and *9* have a sequence of PAMFATAGRK and the polylinkers preceding subunits 3 and 5 in *Lhca1* have a sequence of VAMFASSGHKD, suggesting that the homogenization events detected within the coding regions of these genes extend to the polylinkers (fig. 1*A* and *B*). This even applies to all the polylinkers preceding LHCII-type I subunits (fig. 1*D*) that represent both category 1 and 2 polyproteins having the sequence P/GAMFAASGRK, with the exception of subunit 2 of *LHCPII* where there is greater sequence variation (GAMLATSGRK).

Evidence for the Creation of a LHC Gene Encoding a Single Subunit via Truncation/Deletion of an LHCI Polyprotein Gene

We discovered 2 cDNAs that encode only a single LHC subunit: *Lhcb4* and *Lhca3*. *Lhcb4* encodes the CP29 protein, a minor PSII-associated antenna, and the mRNA size was confirmed by Northern hybridization to be 1.4 kb (fig. 2). *Lhca3* encodes a single LHCI subunit that closely resembles subunit 2 of the category 1 LHCI polyprotein-coding gene *Lhca1* (figs. 1*C* and 5). Following the coding region of subunit 2 is a residual polylinker plus the first 25 amino acids of the N-terminal end of subunit 3, which is truncated by a premature stop codon. Interestingly, the similarity beyond the polylinker ends 8 amino acids before the stop codon, where there was a deletion leading to a frameshift and the introduction of a stop codon. The similarity between *Lhca1* and *Lhca3* was also obvious in the signal sequence/transit peptide (data not shown). There are 2 features that suggest the generation of *Lhca3* was more complex than the introduction of a stop codon by a nonsense mutation, which would lead to the truncation a polyprotein-coding gene. The first is that though there is clear evidence for a conserved targeting domain and subunit 2, it appears as though subunit 1 has been excluded in the process of generating the transcript. The second feature

FIG. 4.—A sequence logo plot of the amino acids in and surrounding the decapeptide linkers of all *Euglena* LHCs.

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

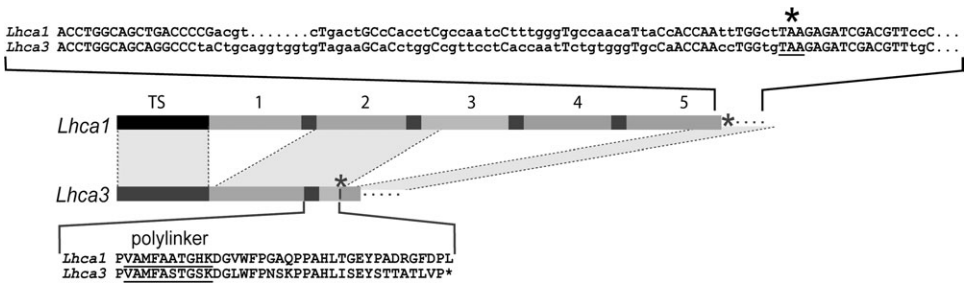

FIG. 5.—Comparison of *Lhca1* and *Lhca3*. Regions of increased sequence identity are linked by dashed lines. The amino acid sequence of the region following subunit 2 for *Lhca1* and *3* is shown (below) as well as the nucleic acid sequence for the 3′UTR of *Lhca3* immediately following its stop codon. Regions from *Lhca1* that correspond to the *Lhca3* 3′UTR are also shown. The dots in the *Lhca1* sequence indicate the presence of a large insertion compared with *Lhca3*. The polylinker region is underlined and the stop codon is indicated with an asterisk.

is the limited similarity to the 3′UTR of *Lhca1* in the sequence following the premature stop codon in *Lhca3* (fig. 5). These findings together suggest that a series of deletions also accompanied the nonsense mutation in creating this LHCI transcript.

Discussion

Polyprotein Organization and Origin

Eukaryotic polyproteins typically contain tandemly repeated coding units that yield mature proteins with highly conserved amino acid sequences. In *Euglena*, however, we observe 3 categories of polyprotein organization: 1) a complex arrangement of divergent subunits, 2) tandem duplications of nearly identical subunits, and 3) tandem duplications of pairs of divergent subunits, though the latter 2 categories are tentative until complete sequencing of the gene family is accomplished. The consistent tandem arrangement of the LHCs in large polyproteins suggests that unequal crossing-over was a driving force in the creation of the multisubunit-encoding genes. This is particularly obvious with the tandem repeats of nearly identical subunits or pairs of divergent subunits as found in categories 2 and 3, respectively. Unequal crossing-over during homologous recombination is a well-described mechanism for the generation of tandemly repeated DNA sequences in genomes (Metzenberg et al. 1991; Propok'ev and Sukhodolets 2005). The creation of multisubunit polyproteins through the tandem duplication of individual subunits has been examined with the *Euglena rbcS* multigene family where, within a polyprotein-coding gene, multiple subunits have conserved intron insertion sites (Tessier et al. 1995), indicating that the subunits are duplicated segments that were incorporated into a single transcriptional unit. Though the partial LHCII genomic clone in GenBank lacked obvious conserved intron insertion sites, examination of the genomic organization of LHC gene family would have to be done to better assess how these large polyproteins were generated.

An obstacle to overcome during the arrangement of the LHC subunits as a polyprotein is the separation of the individual units prior to antenna assembly. In the polyproteins of *Euglena*, the individual coding units are delineated by polylinkers, of which there are 2 major categories: the decapeptide linkers observed in the LHCs and RuBisCO (Chan et al. 1990), and the tetrapeptide linkers with the consensus sequence "SVAM" described for the chloroplast phosphoglycerate kinase (PGK) polyproteins (Nowitzki et al. 2004). Although the presence of different recognition sequences would imply different proteases, the PGK polylinker is composed of amino acids that are chemically similar to the first portion of the LHC polylinkers as well as the 2 amino acids preceding the polylinkers (fig. 4: positions 1–4). In fact, an exact SVAM motif is found in the polylinkers of several LHCI proteins and is apparent in figure 4. Thus, despite containing shorter polylinkers, PGK may be processed by the same protease as the LHCs. N-terminal sequencing of *Euglena* LHCs would be useful in assessing exact processing sites.

The importance of gene duplications in the evolution of multigene families is well known (Hughes 1994). As with the LHC gene family in most eukaryotes, gene duplication is an important evolutionary mechanism in *Euglena*; in addition to the tandem duplications giving rise to the polyproteins, we also found evidence for the duplication of entire polyprotein-coding genes. The 7 category 1 LHCI polyproteins, for instance, are all paralogous as they possess a conserved order of LHC subunits (fig. 1*A*). Similar evidence for gene duplications was found for the LHCII category 2 polyproteins *Lhcbm4*/*Lhcbm8* and *Lhcbm6*/*Lhcbm9*. Gene duplication creates functional redundancies, potentially allowing for the relaxed selection of one copy, that may ultimately lead to sub- or neofunctionalization of the redundant copy (Ohta 1987; Hurles 2004). Such subfunctionalization of LHC genes is well known for the plant and green algal antenna systems where different LHCs have specific interactions and functions within the photosystems (Jansson 1999). In fact, there is a comparable degree of divergence between the subunits within the category 1 polyproteins as found within the different plant/green algal LHCII proteins (ca. 50% amino acid identity). Thus, it is likely that the different subunits of the polyprotein precursors have equivalently complex functional roles in fine-tuning the light-harvesting capacity of the antenna systems.

Intriguingly, some subunits within polyproteins share higher sequence identity with each other than they do with subunits in paralogous polyproteins, indicating that there are homogenizing events acting on these subunits. Many multigene families (Archibald and Roger 2002; Pride and Blaser 2002; Bethke et al. 2006), including those that

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

code for polyproteins (Sharp and Li 1987; Keeling and Doolittle 1995), are subject to sequence homogenization. This is possibly due to unequal crossing-over or biased gene conversion (Hillis et al. 1991), where gene conversion involves the nonreciprocal transfer of genetic information between highly similar sequences during homologous recombination (Abdulkarim and Hughes 1996). The repeated units of a gene family undergoing gene conversion events would evolve in concert with each other, accumulating fewer synonymous and nonsynonymous changes compared with paralogous genes in closely related organisms. The location of homologous genes within the genome appears to affect the level of homogenization as gene conversion occurs at increased frequency in genes closely arranged in head-to-head or tandem configurations (Benedict et al. 1996; Liao 1999). We speculate that the LHCI subunits 3 and 5, within the category 1 LHCI polyproteins are evolving in concert, as is LHCII subunit type I in *Lhcbm1*, *Lhcbm4*, and *Lhcbm8*. The subunits in the category 2 polyproteins *Lhca9*, *Lhcbm6*, *Lhcbm9*, and *Lhcbm10* also appear to be evolving in concert at the intraprotein level. All homogenized subunits contain fewer synonymous changes between the subunits at the intrapolyprotein level than at the interprotein level (data not shown) and are predicted to have undergone gene conversion events (Sawyer 1999). The similarities between the type I subunit of *Lhcbm4*, *Lhcbm8* (category 2 polyproteins), and *Lhcbm1* (a category 1 LHCII polyprotein) are particularly interesting as it is possible that *Lhcbm4* and *Lhcbm8* are acting as donors for recombination, as is suggested for polyubiquitin (Catic and Ploegh 2005), which would allow the homogenization of subunit type I across all polyproteins possessing this subunit. There is also the possibility that the conservation is due to either purifying selection of an essential LHC subunit (Nei and Rooney 2005), as is proposed for the histone gene family (Piontkivska et al. 2002; Eirin-Lopez et al. 2004). However, the trend of highly conserved polylinkers preceding the conserved subunits would imply a homogenization as the polylinker sequence is expected to tolerate significant changes in primary sequence. In addition, LHCs are usually quite divergent in loop regions and a high level of selection in these regions would be unexpected. Examination of the chromosomal location and genomic sequence of these polyprotein-coding genes will help determine the exact extent of gene conversion or other mechanisms driving LHC evolution in *Euglena*.

Not All LHCs Are Encoded by Polyprotein Genes

It is interesting that there are 2 transcripts that do not encode polyproteins: *Lhcb4* (CP29) (Koziol et al. 2007) and *Lhca3*. There is no evidence to suggest that *Lhcb4* mRNA has ever encoded a polyprotein as it contains only a single-coding unit and lacks evidence of polylinkers. *Lhca3*, however, was generated through the rearrangement of a category 1 LHCI polyprotein-coding gene, through a series of internal deletions and the generation of a premature stop codon that resulted in a complete subunit connected to a truncated subunit by a polylinker (fig. 5). As this polylinker is conserved, the protein is likely proteolytically cleaved prior to its insertion into the membrane (Sulli and Schwartzbach 1996). It is possible that the *Lhcb4* gene was generated in a similar manner, but through an ancient series of deletions, the evidence for which is no longer recognizable.

It is unknown why *Euglena* has polyprotein-coding genes for several gene families, though there are a few possible explanations. Initially, the arrangement of LHC-coding units into a polyprotein may have been to increase gene dosage in an organism with a reduced dependence of transcriptional regulatory mechanisms and a reliance on transsplicing to generate translatable mRNAs (McCarthy and Schwartzbach 1984; Keller et al. 1992), as proposed for the tandemly arranged genes in trypanosomes (Jackson 2007). Such an arrangement would allow for the production of a large number of protein subunits whose synthesis could be efficiently controlled by posttranscriptional mechanisms. This may have been linked to another rationale for polyprotein maintenance, one that is usually associated with viral production. The presence of variable numbers of distinct LHC types in polyproteins may assist in maintaining the proper stoichiometry of the different antenna proteins within the thylakoid membranes and to allow for the fine-tuning of their light-harvesting antennae. Though this is more likely to apply to the PSI antennae given that the category 1 polyproteins are predominant and encode for the distinct LHCI subunits that may make up the LHCI antenna belt. Nevertheless, any advantage of such an organization would have to be weighed against the disadvantages of polyprotein-coding genes. For instance, the introduction of frameshift or nonsense mutations in the upstream portion of a gene would deactivate multiple downstream-coding subunits, as we witnessed with *Lhca3*, and yielding pseudogenes (Catic and Ploegh 2005).

For both LHCI- and LHCII-coding genes, up to 17 distinct types were resolved by phylogenetic analyses and most encoded as polyprotein precursors. As it is generally accepted that *Euglena* acquired a plastid secondarily from a green alga and because no green alga described to date have antenna proteins organized into polyproteins, it is probable that the polyprotein conformation was created during the flood of green algal genes into the euglenoid nucleus during the endosymbiotic origin of the plastid. It is particularly interesting that dinoflagellates have independently evolved the use of polyproteins during the secondary origin of their plastids. Though the reason for this organization remains speculative, the large number of unique coding units generated in the process translate into a complex network of LHC subunits. This likely allowed for the fine-tuning of the light-harvesting apparatuses as different paralogs assumed specialized roles within the antenna systems and facilitated the origin of novel light-harvesting and photoprotective strategies.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

Acknowledgments

We thank M.W. Gray for access to *Euglena* EST data, G.W. Saunders for helpful discussions, and D. Clark for reviewing an earlier version of this manuscript. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada, Genome Atlantic, and Genome Canada.

Literature Cited

Abdulkarim F, Hughes D. 1996. Homologous recombination between the tuf genes of *salmonella typhimurium*. J Mol Biol. 260:506–522.

Akaike H. 1974. New Look at Statistical-Model Identification. IEEE Trans Autom Control. 19:716–723.

Archibald JM, Roger AJ. 2002. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J Mol Biol. 316:1041–1050.

Arribas C, Sampedro J, Izquierdo M. 1986. The ubiquitin genes in *Drosophila melanogaster*—transcription and polymorphism. Biochim Biophys Acta. 868:119–127.

Benedict MQ, Levine BJ, Ke ZX, Cockburn AF, Seawright JA. 1996. Precise limitation of concerted evolution to ORFs in mosquito Hsp82 genes. Insect Mol Biol. 5:73–79.

Bethke LL, Zilversmit M, Nielsen K, Daily J, Volkman SK, Ndiaye D, Lozovsky ER, Hartl DL, Wirth DF. 2006. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. Mol Biochem Parasitol. 150:10–24.

Carrington JC, Dougherty WG. 1987. Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. J Virol. 61:2540–2548.

Catic A, Ploegh HL. 2005. Ubiquitin–conserved protein or selfish gene? Trends Biochem Sci. 30:600–604.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol. 46:347–366.

Chan RL, Keller M, Canaday J, Weil JH, Imbault P. 1990. Eight small subunits of *Euglena* ribulose 1-5 bisphosphate carboxylase/oxygenase are translated from a large mRNA as a polyprotein. EMBO J. 9:333–338.

Chen LB, DeVries AL, Cheng CHC. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci U S A. 94:3811–3816.

Comeron JM. 1999. K-estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics. 15:763–764.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14: 1188–1190.

Durnford DG, Gray MW. 2006. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. Eukaryot Cell. 5:2079–2091.

Durnford DG, Price JA, McKim SM, Sarchfield ML. 2003. Light-harvesting complex gene expression is controlled by both transcriptional and post-transcriptional mechanisms during photoacclimation in *Chlamydomonas reinhardtii*. Physiol Plant. 118:193–205.

Eirin-Lopez JM, Gonzalez-Tizon AM, Martinez A, Mendez J. 2004. Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of *orphon* H1 genes. Mol Biol Evol. 21:1992–2003.

Gan SQ, McBride OW, Idler WW, Markova N, Steinert PM. 1990. Organization, structure, and polymorphisms of the human profilaggrin gene. Biochemistry. 29:9432–9440.

Graham RW, Jones D, Candido EP. 1989. UbiA, the major polyubiquitin locus in *Caenorhabditis elegans*, has unusual structural features and is constitutively expressed. Mol Cell Biol. 9:268–277.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33:W557–W559.

Hiller RG, Wrench PM, Sharples FP. 1995. The light-harvesting chlorophyll a-c-binding protein of dinoflagellates: a putative polyprotein. FEBS Lett. 363:175–178.

Hillis DM, Moritz C, Porter CA, Baker RJ. 1991. Evidence for biased gene conversion in concerted evolution of ribosomal DNA. Science. 251:308–310.

Houlné G, Schantz R. 1987. Molecular analysis of the transcripts encoding the light-harvesting chlorophyll a/b protein in *Euglena gracilis*: unusual size of the mRNA. Curr Genet. 12: 611–616.

Houlné G, Schantz R. 1988. Characterization of cDNA sequences for LhcI apoproteins in *Euglena gracilis*—the messenger-RNA encodes a large precursor containing several consecutive divergent polypeptides. Mol Gen Genet. 213: 479–486.

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

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 256:119–124.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2:E206.

Jackson AP. 2007. Origins of amino acid transporter loci in trypanosomatid parasites. BMC Evol Biol. 7:26.

Jansson S. 1999. A guide to the lhc genes and their relatives in Arabidopsis. Trends Plant Sci. 4:236–240.

Keeling PJ, Doolittle WF. 1995. Concerted evolution in protists: recent homogenization of a polyubiquitin gene in *Trichomonas vaginalis*. J Mol Evol. 41:556–562.

Keller M, Tessier LH, Chan RL, Weil JH, Imbault P. 1992. In *Euglena*, spliced-leader RNA (SL-RNA) and 5S rRNA genes are tandemly repeated. Nucleic Acids Res. 20:1711–1715.

Kennedy MW. 2000. The polyprotein lipid binding proteins of nematodes. Biochim Biophys Acta. 1476:149–164.

Koziol AG, Borza T, Ishida KI, Keeling P, Lee RW, Durnford DG. 2007. Tracing the evolution of the light-harvesting antennae in chlorophyll a/b-containing organisms. Plant Physiol. 143:1802–1816.

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet. 64:24–30.

McCarthy SA, Schwartzbach SD. 1984. Absence of photoregulation of abundant messenger-RNA levels in *euglena*. Plant Sci Lett. 35:61–66.

Metzenberg AB, Wurzer G, Huisman TH, Smithies O. 1991. Homology requirements for unequal crossing over in humans. Genetics. 128:143–161.

Muchhal US, Schwartzbach SD. 1992. Characterization of a *Euglena* gene encoding a polyprotein precursor to the light-harvesting chlorophyll a/b-binding protein of photosystem II. Plant Mol Biol. 18:287–299.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39: 121–152.

Nowitzki U, Gelius-Dietrich G, Schwieger M, Henze K, Martin W. 2004. Chloroplast phosphoglycerate kinase from

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024

Euglena gracilis: endosymbiotic gene replacement going against the tide. Eur J Biochem. 271:4123–4131.
Nylander JAA. 2004. MrModeltest v2. [Internet]. Evolutionary Biology Centre, Uppsala University; Available from: http://www.abc.se/~nylander/.
Ohta T. 1987. Simulating evolution by gene duplication. Genetics. 115:207–213.
Osafune T, Schiff JA, Hase E. 1990. Immunogold localization of Lhcp-II apoprotein in the golgi of euglena. Cell Struct Funct. 15:99–105.
Ozkaynak E, Finley D, Varshavsky A. 1984. The yeast ubiquitin gene: head-to-tail repeats encoding a polyubiquitin precursor protein. Nature. 312:663–666.
Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 12:357–358.
Piontkivska H, Rooney AP, Nei M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol Biol Evol. 19:689–697.
Pride DT, Blaser MJ. 2002. Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. J Mol Biol. 316:629–642.
Propok'ev VV, Sukhodolets VV. 2005. Unequal crossing over is the principal pathway of homologous recombination in tandem duplications of *Escherichia coli*. Genetika. 41:1038–1044.
Rowan R, Whitney SM, Fowler A, Yellowlees D. 1996. Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. Plant Cell. 8:539–553.
Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol Biol Evol. 6:526–538.
Sawyer SA. 1999. GENECONV: a computer package for the statistical detection of gene conversion. [Internet]. Department of Mathematics, Washington University in St. Louis; Available from: http://www.math.wustl.edu/~sawyer.
Sharp PM, Li WH. 1987. Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. J Mol Evol. 25:58–64.
Sulli C, Schwartzbach SD. 1996. A soluble protein is imported into euglena chloroplasts as a membrane-bound precursor. Plant Cell. 8:43–53.
Tessier LH, Paulus F, Keller M, Vial C, Imbault P. 1995. Structure and expression of *Euglena gracilis* nuclear *rbcS* genes encoding the small subunits of the ribulose 1,5-bisphosphate carboxylase/oxygenase: a novel splicing process for unusual intervening sequences? J Mol Biol. 245:22–33.

Geoffrey McFadden, Associate Editor

Accepted October 11, 2007

Downloaded from https://academic.oup.com/mbe/article/25/1/92/1093508 by guest on 25 April 2024