PNAS
www.pnas.org

# Supplementary Information for

Polycistronic gene expression is widespread in green algae

Sean D. Gallaher*, Rory J. Craig, Iniyan Ganesan, Samuel O. Purvine, Sean R. McCorkle, Jane Grimwood, Daniela Strenkert, Lital Davidi, Melissa S. Roth, Tim L. Jeffers, Mary S. Lipton, Krishna K. Niyogi, Jeremy Schmutz, Steven M. Theg, Crysten E. Blaby-Haas, Sabeeha S. Merchant**

* Sean D. Gallaher, email: gallaher@chem.ucla.edu

** Sabeeha S. Merchant, email: sabeeha@berkeley.edu

**This PDF file includes:**

Supplementary text

Figures S1 to S10

Legends for Datasets S1 to S9

SI References

**Other supplementary materials for this manuscript include the following:**

**Dataset S1.** Master Table of Polycistronic Loci

**Dataset S2.** Proteomics Analysis

**Dataset S3.** Sequences of Constructs for *in vitro* Transcription and Translation

**Dataset S4.** Homology Analysis

**Dataset S5.** Functional Annotations

**Dataset S6.** PredAlgo Analysis

**Dataset S7.** IRESfinder Analysis

**Dataset S8.** Annotations for *C. reinhardtii* (v5.7)

**Dataset S9.** Annotations for *C. zofingiensis* (v5.3)

## *Supplementary Results*

### The polycistronic loci are unlikely to be misidentified selenoproteins.

In genes that encode selenoproteins, the UGA stop codon is repurposed to code for a selenocysteine. As such, the selenocysteine codon is typically mis-identified as a stop codon by commonly used gene prediction tools. Given that the *C. reinhardtii* genome encodes selenoproteins (1), this raised the possibility that some of the polycistronic transcripts identified in this work are in fact uncharacterized selenoproteins. For polycistronic loci to be mischaracterized selenoproteins genes, 1) the upstream and downstream ORFs would have to be in the same reading frame, and 2) the two ORFs would have exclusively UGA stop codons between them. Only 11 out of 87 polycistronic loci in *C. reinhardtii*, and 15 out of 173 polycistronic loci in *C. zofingiensis* met both criteria. However, SECISearch3 (2), which identifies a distinctive stem loop structure in selenoprotein transcripts called the SECIS element, failed to identify such an element in any of these 26 loci. Furthermore, none of the ORFs in the 26 loci showed significant sequence similarity to any known selenoproteins when evaluated by Seblastian (3). On the other hand, seven monocistronic *C. zofingiensis* genes were identified with SECIS elements and homology to known proteins. We conclude that none of the UGA-containing polycistronic loci corresponds to a selenoprotein.

### Cellular Targeting of Polycistronically-expressed Proteins

When PredAlgo (4) was used to predict the intracellular targets of all proteins in the curated polycistronic set, we identified examples of organellar targeted pairs: two to the chloroplast (Cre12.g486850/Cre12.g486851 and Cre16.g658245/Cre16.g658250) and four to the mitochondrion (Cz03g17075/Cz03g17085, Cz07g15200/Cz07g15210, Cz09g18275/Cz09g18280, Cz10g12030/Cz10g12040). However, the percentage of polycistronically expressed proteins predicted to be targeted to the mitochondria (*C. reinhardtii* = 19%, *C. zofingiensis* = 11%) or the chloroplast (*C. reinhardtii* = 14%, *C.*

*zofingiensis* 11%) was not significantly different than the percentages for all proteins (Dataset S6).

### Do polycistronic inter-ORF sequences function as IRESs?

ORFs may be translated from the traditional 5'-cap mediated method or by 5'-cap independent means; namely IRESs. The discovery of polycistronic expression in several species of chlorophyte algae led to the question: is translation of downstream ORFs in polycistronic transcripts mediated by IRESs in the inter-ORF sequences?

To systematically evaluate the role of IRESs in polycistronic loci, the inter-ORF sequences of all polycistronic transcripts from *C. reinhardtii* and *C. zofingiensis* were subjected to analysis by the computational IRES prediction tool, IRESfinder (5). For this analysis, 87 inter-ORF sequences were isolated from the polycistronic loci of *C. reinhardtii,* and 204 inter-ORF sequences were isolated from *C. zofingiensis*. For comparison, size-matched sequences selected randomly from the intergenic regions of both species were also analyzed. As controls, 1000 sequences empirically determined to function as IRESs (the "IRES" set) and 1000 sequences empirically determined not to function as IRESs (the "non IRES" set) were also analyzed. These sequences and their IRESfinder scores are presented in Dataset S7 and summarized in Figure S10.

The distribution of IRESfinder scores for the inter-ORF sequences of polycistronic loci in *C. reinhardtii* and *C. zofingiensis* were significantly lower than the scores for known IRESs. The distribution of scores was not significantly different for the polycistronic inter-ORF sequences relative to matched sets of random sequence. While these findings do not rule out the possibility that any one or a few polycistronic inter-ORF sequence could facilitate cap-independent translation, they do suggest that cap-independent translation is not the main driver of expression for downstream ORFs in polycistronic loci.

### Polycistronic loci may arise from fragmentation of a larger gene

For a few of the polycistronic loci, we observed that each of the ORFs has a shared domain with a larger, monocistronically expressed gene. This pattern is consistent with a gene duplication, followed by a fragmentation of the large ORF into two or more smaller ORFs. An example of this from a tricistronic locus in *C. reinhardtii* (Cre07.g317438/Cre07.g317421/Cre07.g317400) is presented in Figure S9.

## *Supplementary Methods*

### Structural gene annotation

In a comparison of *C. reinhardtii* with other species of Chlamydomonas (6), nearly 100 genes were identified that are not present in the current annotations (v5.6, phytozome.net). Of these unannotated genes, 25 genes were found to be consistently co-translated with annotated genes. These 25 genes were manually added to the v5.6 gene annotations to produce v5.7 (Dataset S8), which was used for this work. Additional edits were made to better reflect the Iso-Seq data. In total, 30 new genes were added, and 96 existing genes were updated. All new and edited gene models can be identified in the annotation file by the "v5.7" tag.

In contrast to *C. reinhardtii*, the current annotations for *C. zofingiensis* were observed to be highly fragmented and inaccurate in comparison to RNA-Seq and Iso-Seq data. *De novo* gene annotation was performed with Braker, using available RNA-Seq data (see below) as input. The resulting annotations were merged with the previous annotations (v5.2.3.2, phytozome.net) and Iso-Seq data to produce the v5.3 annotations that were used in this work (Dataset S9).

### uORFs

All uORFs were identified in the transcriptomes of *C. reinhardtii* and *C. zofingiensis* with in-house scripts. Briefly, each transcript and each coding sequence (i.e. primary ORF) were compared to identify the 5' UTR. This sequence was scanned in all three forward open reading frames for sequences between a start codon and a stop codon. For uORFs with multiple start codons, only the largest uORF was considered. Next, the uORFs were classified as class 1 (in frame with the primary start codon, thus extending the primary ORF), class 2 (an ORF that overlaps, but is out of frame with the primary start codon), and class 3 (an ORF that initiates and terminates entirely within the 5' UTR) according to the scheme used by Cross (7). For uORF size, only class 2 and class 3 uORFs of 9 nt or larger were considered. For uORF spacing, only class 3 uORFs of 9 nt or larger were considered.

## Characterization of polycistronic and monocistronic loci

Stop codon usage was determined from the last three nucleotides of each ORF for every gene with in-house scripts. The ORFs were characterized as "polycistronic upstream", "polycistronic downstream", or "monocistronic" (all remaining genes), and the distribution of each was plotted in R. Over- and under-representation of stop codons was calculated using the upper cumulative and lower cumulative probability from the hypergeometric distribution. Pairs of polycistronic loci were scored as being in frame (inter-ORF sequence divisible by three) or not by manual examination of each locus on IGV. The results were plotted in R. ORF size and inter-ORF spacing were calculated from the gff3-formated annotation files for *C. reinhardtii* (v5.7) and *C. zofingiensis* (v5.3) with in-house scripts, and plotted in R.

## H3K4me3 ChIP-Seq in *C. reinhardtii*

ChIP was performed as described previously (8) with the following modifications: *C. reinhardtii* (strain CC-5390) was grown in a photo-bioreactor as described previously (9). A total of $5x10^7$ cells, corresponding to 50 mL culture, were collected by a 2 min centrifugation at 4°C and 3220 g. Supernatant was discarded completely. To cross-link protein–DNA interactions, cells were resuspended in 10 mL freshly prepared cross-linking buffer (20 mM HEPES-KOH, pH 7.6, 80 mM KCl, and 0.35% formaldehyde) and incubated for 10 min at 24°C. Cross-linking was quenched by the addition of Glycine to a final concentration of 125 mM and incubation for 5 min at 24°C. Cells were collected by a 2 min centrifugation at 4°C and 3220 g. Cells were lysed by the addition of 1000 µL lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.0, and 0.25× protease inhibitor cocktail [Roche]). Cells were sonicated on ice to achieve an average DNA fragment size of ~200 nt. ChIP was performed with aliquots corresponding to ~1 × $10^7$ cells that were diluted 10-fold with ChIP buffer (1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, and 16.7 mM Tris-HCl, pH 8) and supplemented with bovine serum albumin. We omitted the addition of sonicated λ-DNA that was supplemented as blocking agent in (8) since it interferes with library preparation steps after ChIP. 10 µL of a ChIP-grade antibody specific for trimethylated H3K4 (Abcam, ab8580) was added, and antibody-protein/DNA complexes were allowed to form during a 1 h incubation at 4°C. Next, samples were complexed with 6 mg preswollen protein A Sepharose beads (Sigma-Aldrich) during a 2 h incubation at 4°C, and precipitated by a 20 s centrifugation at 16,000 g. Sepharose beads were washed once with washing buffer 1 (0.1% SDS, 1% Triton X-100, and 2 mM EDTA, pH 8) containing 150 mM NaCl, once with washing buffer 1 containing 500 mM NaCl, once with washing buffer 2 (250 mM LiCl, 1% Nonidet P-40, 1% Na-deoxycholate, 1 mM EDTA, and 10 mM Tris-HCl, pH 8), and twice with TE (1 mM EDTA and 10 mM Tris-HCl, pH 8). Protein-DNA complexes were eluted by incubating twice for 15 min at 65°C in elution buffer (1% SDS and 0.1 M $NaHCO_3$), and cross-links were reverted by an overnight incubation at 65°C after addition of NaCl to a final concentration of 0.5 M. Proteins were digested by incubating for 1 h at 55°C after the addition of proteinase K (3.5 µg/mL), EDTA (8 mM), and Tris-HCl, pH 8.0 (32 mM). DNA was extracted once with phenol/chloroform/isoamyl alcohol (25:24:1), once with chloroform/isoamyl alcohol (24:1), and precipitated by incubation with 2 volumes of ethanol after addition of 0.3 M Na-acetate, pH 5.2, and 10 µg/mL glycogen for 3 h at −20°C. Precipitated DNA was collected by a 20 min centrifugation at 4°C and 16,000 g, washed with 70% ethanol, air-dried, and resuspended in 50 µL MilliQ water. Library preparation on immunoprecipitated DNA was performed using the TruSeq ChIP Library Preparation Kit according to the manufacturer’s instructions (Illumina). qPCR was used to determine the concentration of the libraries. Libraries were sequenced on a HiSeq

2000 sequencer (Illumina), and the data resulting from 12 samples was pooled and mapped to the *C. reinhardtii* genome assembly (v5, Phytozome.net) with bwa mem (10). Peaks of H3K4me3 enrichment were identified with MACS (11). A score was calculated from the -$\log_{10}$ of the FDR-adjusted *p*-value for each locus in the genome, and was formatted as a bedgraph file for viewing on IGV. The mean score for the first 500 nt of each gene was calculated with in-house scripts, and plotted in R.

## Polyadenylation signal analysis

The most frequently used polyadenylation signal in *C. reinhardtii*, UGUAA, was identified previously (12, 13). For *C. zofingiensis*, the portion of Iso-Seq reads from –25 to –5 nt upstream of a string of poly(A) was isolated with in-house scripts. A *k*-mer analysis was performed on these sequences with Jellyfish (14), and the same 5-mer, UGUAA, was found to be significantly over-represented. The final 100 nt of each computationally-annotated transcript model (i.e.,one monocistronic transcript per ORF) were isolated with in-house scripts. These sequences were scored for the presence of "UGUAA", and the percentage of genes with this sequence was plotted in R. The likelihood of a "UGUAA" to occur in randomly generated 100 nt sequences was calculated at 3.2% for *C. reinhardtii*, and 8.8% for *C. zofingiensis* based on 64% and 51% GC content, respectively (15, 16).

## Poly(A) tail quantification

The 100 nt immediately upstream of a stretch of ≥8 As was isolated from the Iso-Seq CCS data from *C. reinhardtii* and *C. zofingiensis* with in-house scripts. These pseudo-reads were mapped to the appropriate genome assembly with minimap2 version 2.17-r941 (17) with the following settings: -x sr -a --frag=no --splice -B4. Full length CCS reads were mapped minimap2 with the following settings: -x splice:hq -a -G 20k. The resulting sam-formatted alignment files were converted to bam, sorted, and indexed with samtools version 1.9-58-gbd1a409 (18). In-house scripts were used to generate bed-formatted tables for each species with the stand and position info of the 3' terminal 1000 nts of each gene, or the full gene for genes <1000 nts. The number of read counts from the poly(A)-adjacent pseudo-reads and from the full length Iso-Seq CCS reads were determined for the 3' terminal 1000 nt intervals for each gene with bedtools multicov version 2.25.0 (19) using the following settings: -s -D. The ratio of poly(A)-adjacent pseudo-reads relative to total Iso-Seq CCS reads was calculated for each gene in which ≥10 Iso-Seq CCS reads were counted. The distribution of these per-gene ratios were determined for polycistronic upstream, polycistronic downstream, and monocistronic (all remaining) loci, and plotted as box plots in R.

## Iso-Seq library construction

Total RNA from several previously described conditions in *C. reinhardtii* (9, 20–24) and *C. zofingiensis* (16) was pooled for library preparation. Full length cDNA was synthesized from 1 ug of total RNA with the SMARTer PCR cDNA Synthesis kit (Clontech). First-strand cDNA was amplified with PrimeSTAR GL DNA Polymerase (Clontech) using template switching oligos to make double-stranded cDNA. Double-stranded cDNA was purified with AMPure PB beads. The amplified cDNA was end-repaired and ligated with blunt end PacBio sequencing adaptors using the SMRTbell Template Prep Kit 1.0 (PacBio). The ligated products were treated by exonuclease to remove unligated products, and then purified again by AMPure PB beads.

## Iso-Seq sequencing

PacBio Sequencing primers were annealed to the SMRTbell template library, and sequencing polymerase was bound to them using Sequel Binding kit 2.0 (PacBio). SMRTbell libraries from *C. reinhardtii* were sequenced on a Sequel II sequencer (PacBio) for 1800 minutes at the HudsonAlpha Institute for Biotechnology. SMRTbell libraries from *C. zofingiensis* were sequenced on a Sequel sequencer (PacBio) using v3 sequencing primer, 1M v2 SMRT cells, and Version 2.1 sequencing chemistry with 600 minute sequencing movie runs at the Joint Genome Institute. The resulting sequencing data was used to produce circular consensus sequences (CCS) using the SMRT Tools suite v3.4 (PacBio). These were aligned to the *C. reinhardtii* (v5, phytozome.net) and *C.*

*zofingiensis* (v5, phytozome.net) genome assemblies with minimap2 (17). The resulting alignment files were visualized with the IGV genome browser (25).

## RNA-Seq

*C. reinhardtii* RNA-Seq data were from a study of diurnal gene expression that compared 16 time points during a 12/12 h light/dark cycle in synchronized cells (9). *C. zofingiensis* RNA-Seq data came from cultures prepared from 14 different growth regimes (Fe, N, and S nutrient limitation, phototrophic growth, heterotrophic growth on glucose, oxidative stress, anaerobic stress, and a range of illumination conditions) designed to capture a broad range of the transcriptome (16). All RNA-Seq data were aligned to the appropriate genome assembly with RNA-Star, and visualized with IGV (25, 26).

## Co-expression analysis

The RNA-Seq data described above was used to quantify transcript abundances for both *C. reinhardtii* and *C. zofingiensis* in terms of Fragments Per Kilobase of transcript per Million mapped reads (FPKMs) with cuffdiff as described previously (9, 16, 27). Weakly expressed genes were excluded from this analysis by filtering genes with maximum FPKMs ≤1. Next, pairs of colinear genes (defined here as genes on the same strand of the same chromosome with ≤20,000 nt between ORFs) were identified for both species with in-house scripts. Colinear genes were sorted into separate lists of polycistronic and monocistronic. The correlation in expression between the first and second gene in each pair of colinear genes was calculated as a PCC with the cor function in R. The distribution of these PCC values was plotted as a box plot for monocistronic and polycistronic colinear genes in R.

## Proteomics

Proteins from *C. reinhardtii* and *C. zofingiensis* were identified by mass-spectrometry from a pool of experiments as described previously (28). Identified peptides were "reverse translated" *in silico* with in-house scripts to produce fasta files of pseudo-cDNA "reads". These were mapped to the appropriate genome assembly and visualized on IGV exactly as the RNA-Seq data. Spectral counts that could not be assigned unambiguously to a single gene were filtered. The total count of the remaining (unambiguously assigned) peptides was determined for each gene, and used to calculate the percentage of polycistronic upstream, polycistronic downstream, and monocistronic (all remaining) genes that had been detected. The same analysis was performed using only N-terminal or only C-terminal peptides. These results were plotted as a bar plot in R. Tables of identified peptides and their corresponding gene IDs are provided in Dataset S2.

### *In vitro* coupled transcription and translation

Sequences of polycistronic loci from both *C. reinhardtii* and *C. zofingiensis* were synthesized and cloned into pET-21(+) expression vectors (Twist Bioscience). Cloned sequences are included in Dataset S3. Additional constructs were designed in which the upstream Kozak-like sequences was made to be more or less favorable to translation initiation, and in which the upstream ORF, downstream ORF, or both, were replaced with the gene encoding mVenus or ribosomal protein RPS14-Em$^R$ (encoded by the *cry1-1* allele) that confers resistance to the drug emetine (29, 30). The resulting expression vectors were subjected to *in vitro* coupled transcription and translation with the TNT wheat germ T7 kit following manufacturer's instructions (Promega). Proteins were radio-labeled during translation with 0.4 mCi/mL [$^{35}$S]-Met. Translations were diluted 10-fold in buffer containing 125 mM Tris–HCl (pH 6.8), 4% SDS, 20% glycerol, 0.05 mg/mL bromophenol blue, and 10% beta-mercaptoethanol. 10 µL of the resulting solution was loaded on Any kD Mini-PROTEAN TGX Precast Protein Gels (Bio-Rad) for separation by electrophoresis for 35 min at 180 V. Protein sizes were estimated against a Precision Plus protein ladder (Bio-Rad). The gels were washed first with glacial acetic acid, then 1 mM 2,5-Diphenyloxazole in glacial acetic acid, and finally DI water. The gels were dried, subjected to fluorography, and quantified as described previously (31). Signal intensities were normalized based on the number of Met in the polypeptide sequence.

## Homology analysis

First, the protein sequences encoded by all ORFs expressed from polycistronic loci in *C. reinhardtii* and *C. zofingiensis* were used as query sequences in a protein-protein similarity search with the DIAMOND protein aligner (32) against the proteomes of the opposite species plus six other chlorophyte species available from Phytozome.net (*Coccomyxa subellipsoidea* C-169 v2.0, *Dunaliella salina* v1.0, *Ostreococcus lucimarinus* v2.0, *Micromonas pusilla* CCMP1545 v3.0, and *Volvox carteri* v2.1). Second, those proteins that did not have a high scoring hit with DIAMOND were subjected to a second round of protein-protein similarity search with BLASTP. Third, in an effort to capture potential unannotated ORFs, query proteins without a high scoring hit from either DIAMOND or BLASTP were used to search the six-frame translated DNA sequences of each species' genome assembly with TBLASTN. All hits from these three searches with a BIT score >30 were loaded into a SQL database for further analysis. This database was then queried to identify pairs of colinear ORFs (ORFs on the same reading strand separated by <10,000 nt) as possible polycistronic loci. This list of candidates was manually curated using the JBrowse tool from Phytozome.net, and filtered to remove pairs of hits to the same ORF. When available, expressed sequence tag (EST) data from Phytozome was used to identify transcripts spanning two or more ORFs in candidate polycistronic loci. Similarly, Iso-Seq data (described above) from *C. reinhardtii*, *C. zofingiensis*, and *D. salina* were used to determine if colinear ORFs are transcribed from polycistronic mRNA in those species. The resulting analysis is included in Dataset S4. A heatmap summarizing the results of this analysis was generated in R.

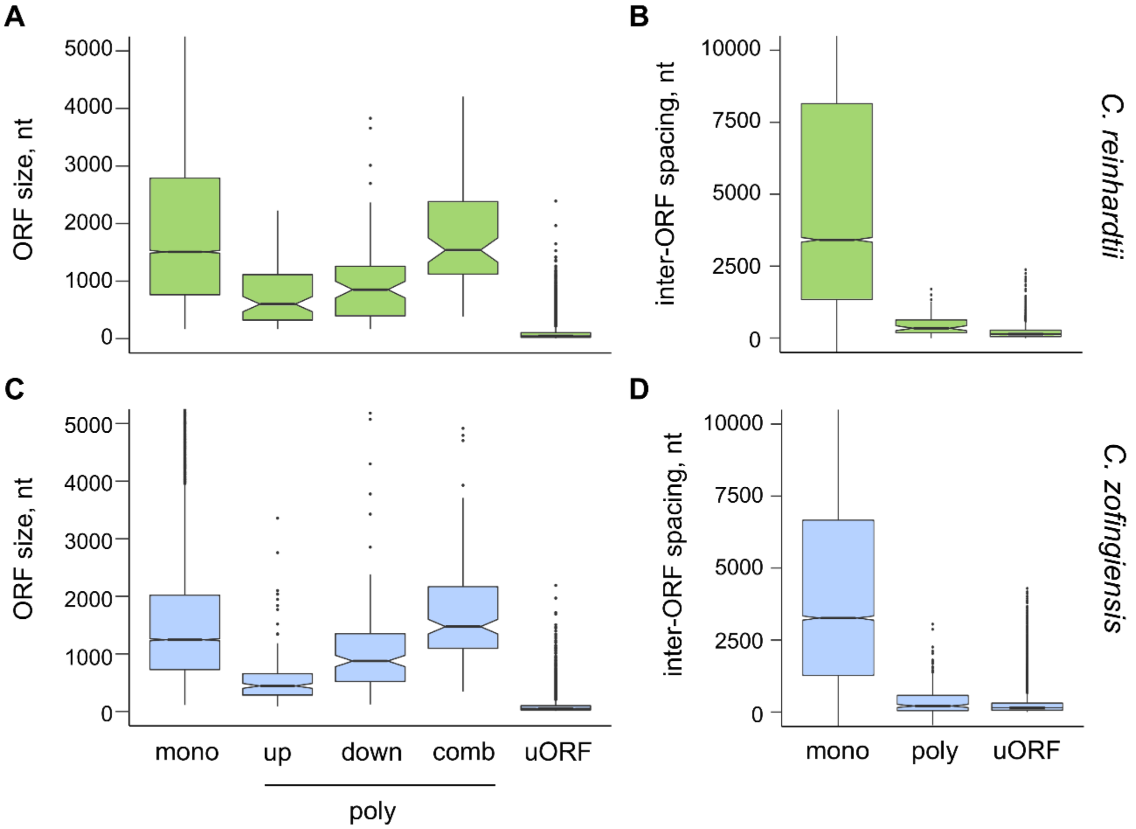
## Functional analysis

The protein sequences encoded by all ORFs expressed from polycistronic loci in *C. reinhardtii* and *C. zofingiensis* were subjected to a search for conserved protein domains in the CDD (v3.18) database with the CD-search tool available from NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) using default settings. The resulting hits were manually curated and are presented in Dataset S5.

## IRESfinder Analysis

The sequences of all inter-ORF regions from the polycistronic loci of *C. reinhardtii* and *C. zofingiensis* were isolated for computational analysis to predict IRES function. For polycistronic loci with three or more ORFs, the sequence between each pair of adjacent ORFs was examined separately. Polycistronic gene pairs with overlapping ORFs (i.e. no inter-ORF sequence) were ignored. A single inter-ORF sequence from *C. zofingiensis* between Cz03g00090 and Cz03g00100 was too small (3 nt) to be analyzed by IRESfinder and was omitted. As a control, random stretches of intergenic sequence from *C. reinhardtii* and *C. zofingiensis* were isolated from both species with in-house scripts. Over 1000 sequences were selected at random from both species and matched for size to the polycistronic inter-ORF sequences. Additional controls for this analysis were selected from work by Weingarten-Gabbay et al. (33). In this work, the authors evaluated 55,000 short nt sequences from a range of sources in a high-throughput, bicistronic reporter assay to quantify the degree to which each sequence can facilitate cap-independent translation. The full set of assayed sequences was filtered exactly as described in Wang and Gribskov (34) to exclude synthetic sequences and to include sequences with "splicing_score" > –2.5 and "promoter_activity" <0.2. From this reduced set, the 1000 sequences with the highest "ires_activity" were selected as the "IRES" set. For the "non IRES" set, 1000 sequences were randomly selected from the remaining 21,367 sequences that exhibited only baseline IRES function ("ires_activity" = 206.29). IRESfinder was downloaded from github (https://github.com/xiaofengsong/IRESfinder, accessed 15-NOV-2020). Each set of sequences described above was subjected to analysis by IRESfinder in mode 0 with default settings. The resulting scores were imported into R and plotted as a box plot with ggplots2 as described in the main text methods. All of the sequences analyzed by IRESfinder and their scores are included in Dataset S7.

## Supplemental Figures S1 - S10



### Figure S1. Structural features of polycistronic loci.

(**A**) The distribution of ORF sizes in nt is plotted as a box plot for *C. reinhardtii* for various classes of ORFs as follows. All monocistronically expressed protein coding ORFs are plotted as "mono" (*N* = 17,594). For polycistronically expressed genes (*N* = 87), the most upstream ORF ("up"), most downstream ORF ("down"), and the sum of all ORF sizes of all co-expressed ORFs on a common transcript ("combined") is plotted. In the final box, the distribution of sizes of all class 2 and class 3 uORFs is plotted ("uORF", *N* = 32,572) (**B**) The inter-ORF spacing between co-linear gene pairs (adjacent genes on the same strand of the same chromosome with ≤ 20,000 nt between ORFs) is plotted for *C. reinhardtii* for monocistronic ("mono", *N* = 15,773) and polycistronic ("poly", *N* = 89) gene pairs. The spacing between all class 3 uORFs and their corresponding primary ORF is plotted as "uORFs" (*N* = 29,484). (**C**) The same analysis as described for panel A was performed on *C. zofingiensis* for monocistronic (*N* = 13,585) and polycistronic (*N* = 173) ORFs and for uORFs (*N* = 49,916). (**D**) The same analysis as described for panel B was performed on *C. zofingiensis* for monocistronic genes (*N* = 13,200), polycistronic (*N* = 215) genes, and uORFs (*N* = 46,700). For each box plot, whiskers indicate 1.5 times the interquartile range and notches indicate the confidence interval of the median. Outliers are plotted as individual points.

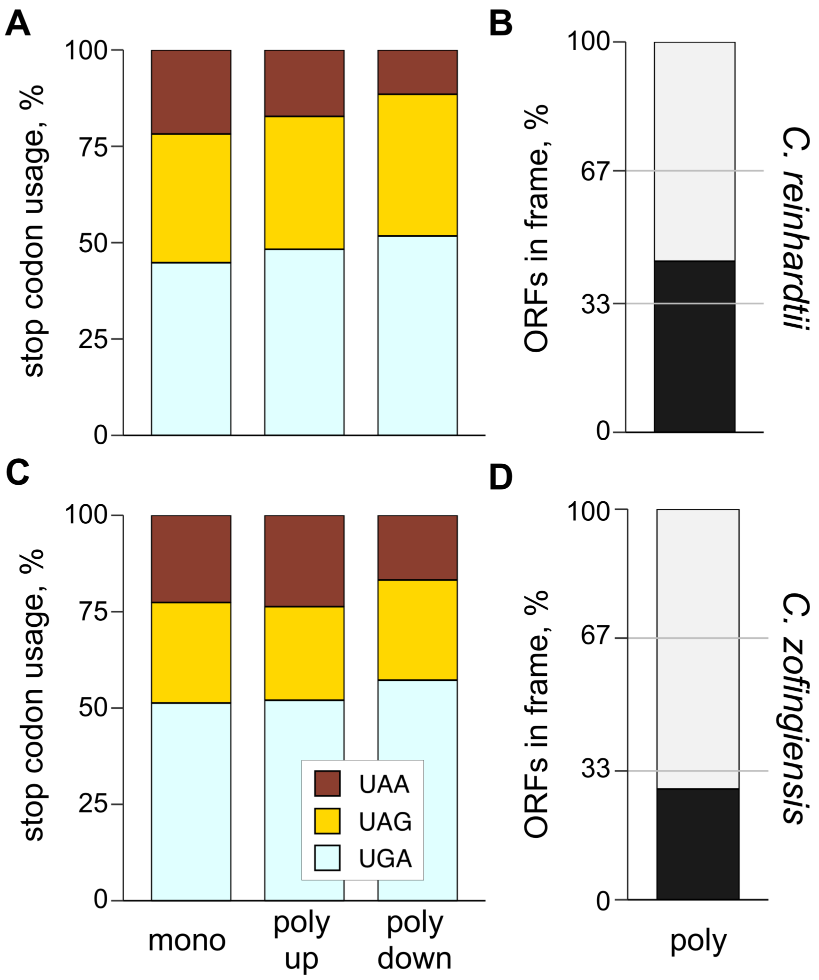


**Figure S2. Frame and stop codon usage in polycistronic loci.**

(**A**)The frequency of stop codon usage in *C. reinhardtii* is plotted for monocistronic ("mono", *N* = 17,570), polycistronic upstream ("poly up", *N* = 87) and polycistronic downstream ("poly down", *N* = 87) genes for. UAA, UAG, and UGA are indicated by the color legend. None of the stop codons is significantly over- or under-represented for the polycistronic upstream and downstream genes ($p$ < 0.05). (**B**) The number of polycistronic gene pairs where both ORFs are in the same reading frame is plotted in black for *C. reinhardtii* (*N* = 89). (**C**) The same analysis as described for panel A was performed on *C. zofingiensis* for monocistronic (*N* = 13,561), polycistronic upstream (*N* = 173), and polycistronic downstream (*N* = 173) genes. (**D**) The same analysis as described for panel B was performed on *C. zofingiensis* for all polycistronic genes (*N* = 215).

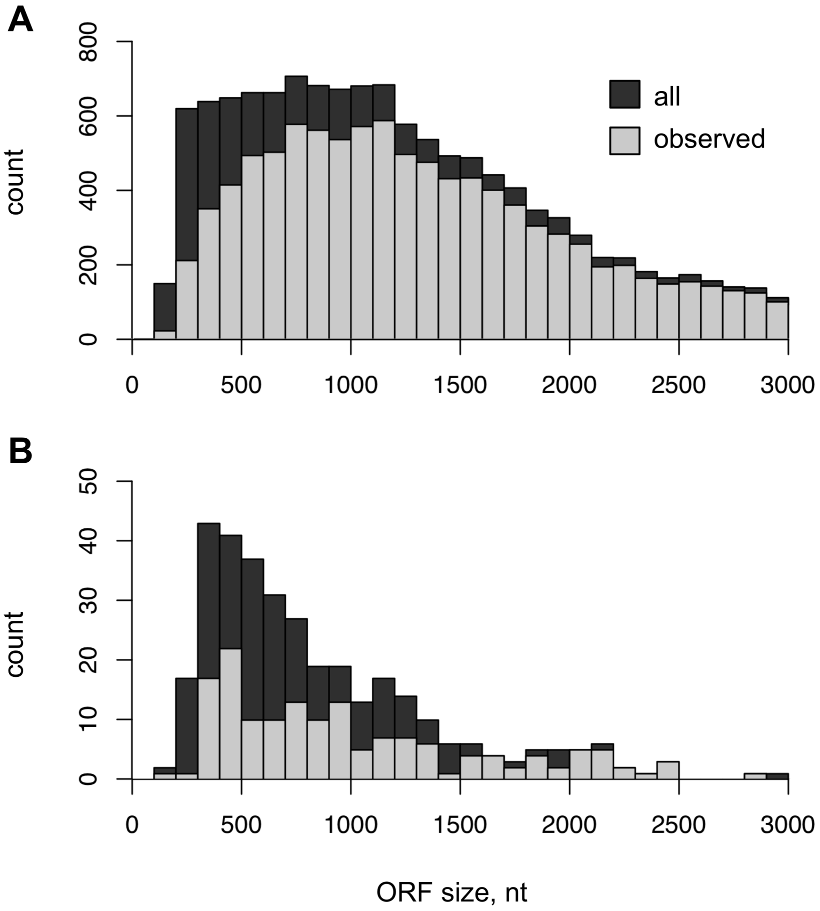
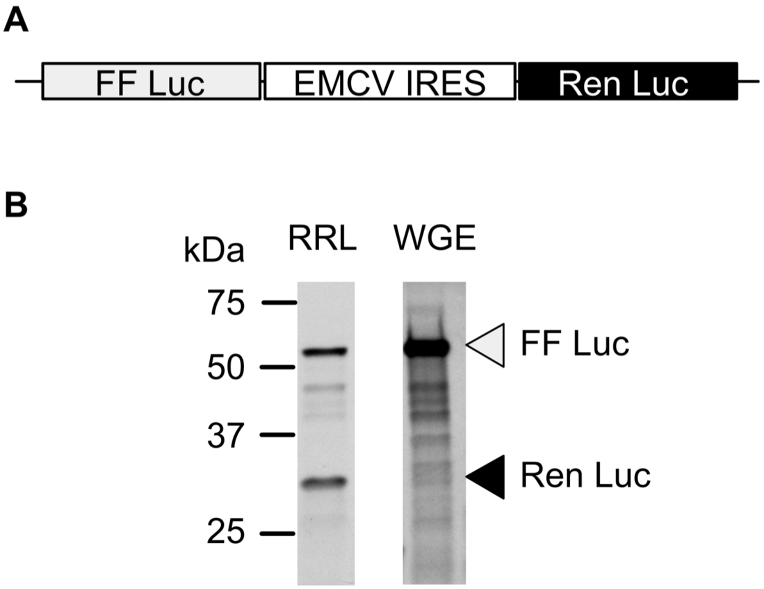


**Figure S3. Detection of proteins as a function of size.**

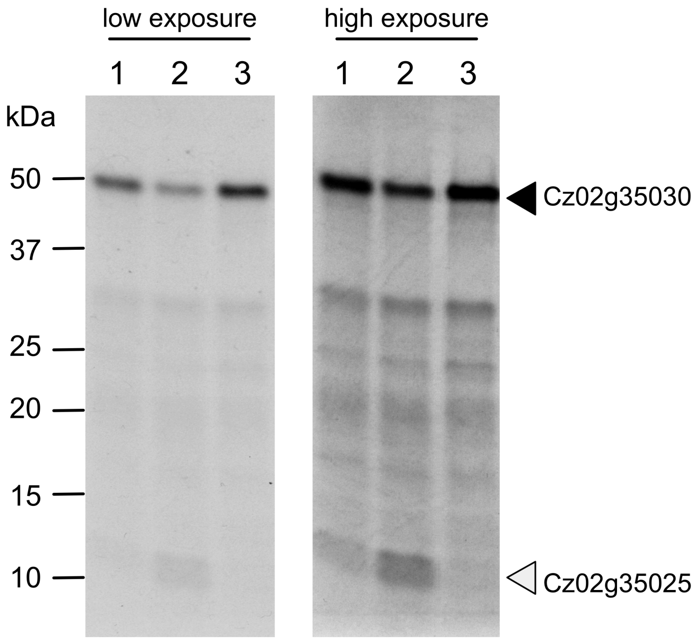
Data from several proteomics studies of *C. zofingiensis* were pooled and analyzed as described in Main Figure 5. Proteins were considered "observed" if there was at least one spectral count of a peptide that could be unambiguously assigned to single protein. (**A**) In the bin histogram presented here, each ORF is sorted based on size and presented as a dark gray bar. ORFs that encode a protein that was observed are presented with light gray bars. For the bottom quartile of ORFs sorted by size, 59.5% were detected in this dataset. For the top quartile, 92.5% of ORFs were detected. (**B**) The same analysis that was described for panel A was performed on the subset of ORFs that are expressed on polycistronic transcripts.



| IVT | Upstream Gene | Size, kDa | Downstream Gene | Size, kDa | Ratio up:down |
|---|---|---|---|---|---|
| RRL | FF Luc | 60.7 | Ren Luc | 36.0 | 0.7 |
| WGE | FF Luc | 60.7 | Ren Luc | 36.0 | 44 |

**Figure S4. EMCV-IRES functions in RRL but not WGE.**

(**A**) A construct was produced that included firefly ("FF Luc") and Renilla ("Ren Luc") luciferases separated by an EMCV IRES. (**B**) This construct transcribed and translated *in vitro* in wheat germ extract (WGE) exactly as described in main text Figure 6. In addition, the same construct was transcribed and translated in rabbit reticulocyte lysate (RRL). The resulting gene products are presented here. The ratio of the upstream gene product (FF Luc) to the downstream gene product (Ren Luc) in both reactions is presented in the accompanying table.



| Lane | Upstream Gene | Size, kDa | Downstream Gene | Size, kDa | Ratio up:down |
|---|---|---|---|---|---|
| 1 | Cz02g35025 | 11.0 | Cz02g35030 | 49.0 | 1.0 |
| 2 | Cz02g35025 | 11.0 | Cz02g35030 | 49.0 | 3.3 |
| 3 | Cz02g35025 | 11.0 | Cz02g35030 | 49.0 | 0.5 |

### Figure S5. Low versus high exposure of Figure 6.

The panel on the right is the same image that is presented in main text Figure 6 (49 d exposure). The panel on the left is the same gel with a lower exposure (17 d exposure). Quantification of the protein bands was performed on unsaturated signals and calibrated against standards.

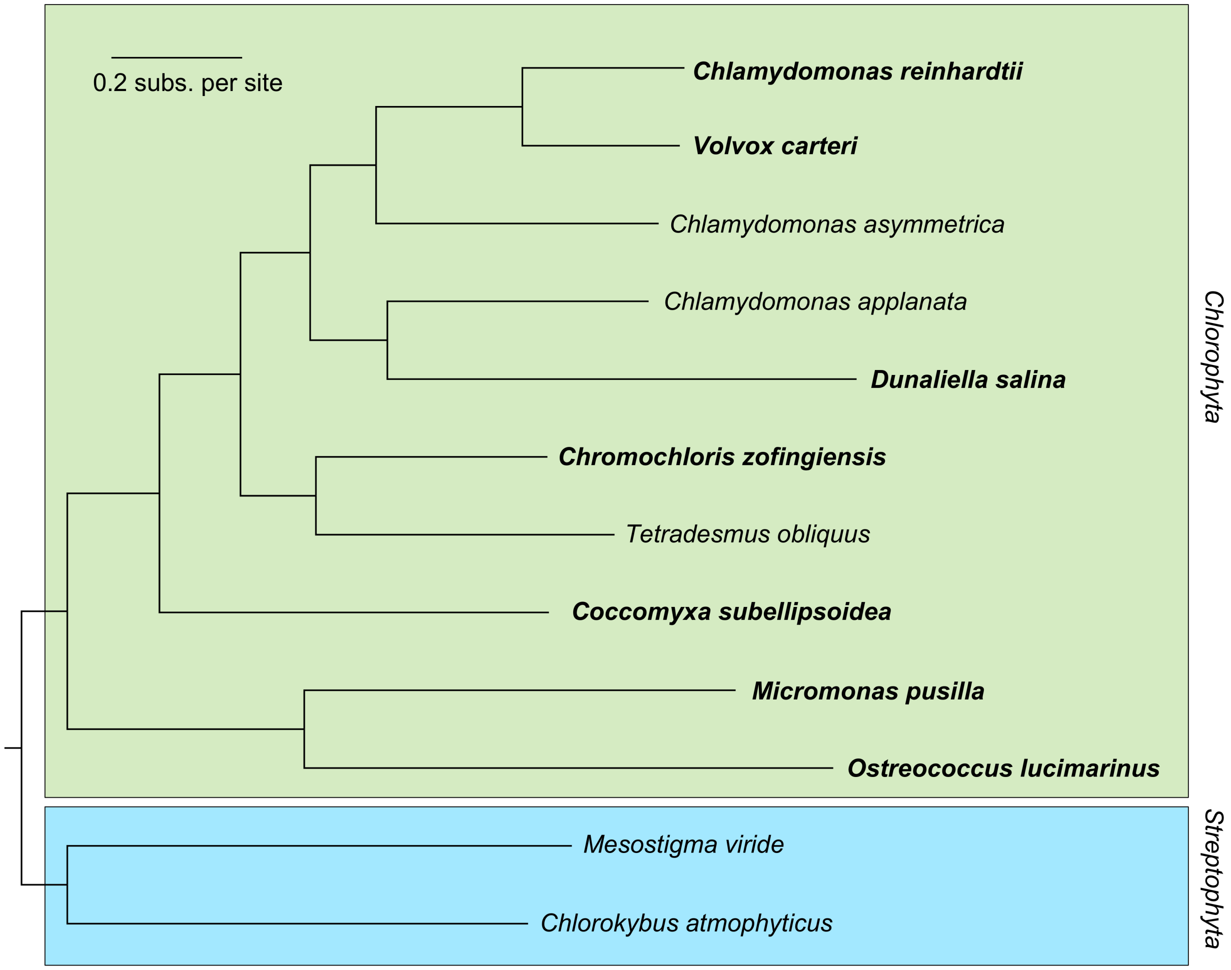


**Figure S6. Phylogenetic tree.**

Concatenated protein sequences from 1,404 single-copy orthologs identified in ten chlorophyte and two streptophyte algal species were used to construct a maximum likelihood phylogenetic tree under the LG+F+R7 model using IQ-TREE (35). The tree is rooted on the Streptophyta clade. Ultrafast bootstrap values were 100 at each branch point. The seven species indicated in bold were used in the analysis of conservation of polycistronic loci that is presented in Main Figure 9.

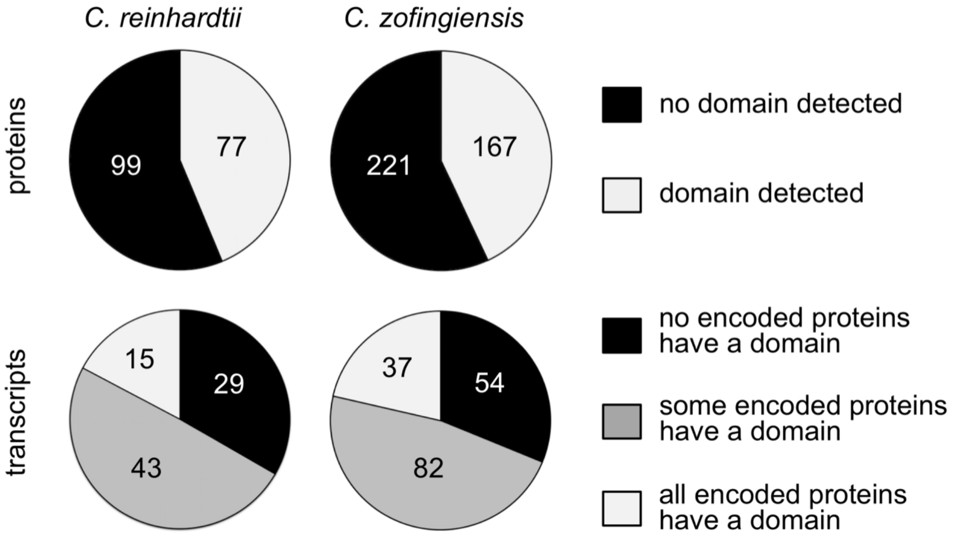


**Figure S7. Conserved domains in polycistronic loci.**

Protein sequences from all polycistronic loci in *C. reinhardtii* and *C. zofingiensis* were searched for conserved domains with the NCBI CD-Search algorithm using default criteria. The number of proteins with an identified domain is presented as a pie chart for both species. Each polycistronic transcript was then scored based on if none, some, or all of the encoded proteins had an identifiable protein domain. The number of polycistronic transcripts in each category is presented as a pie chart for both species.

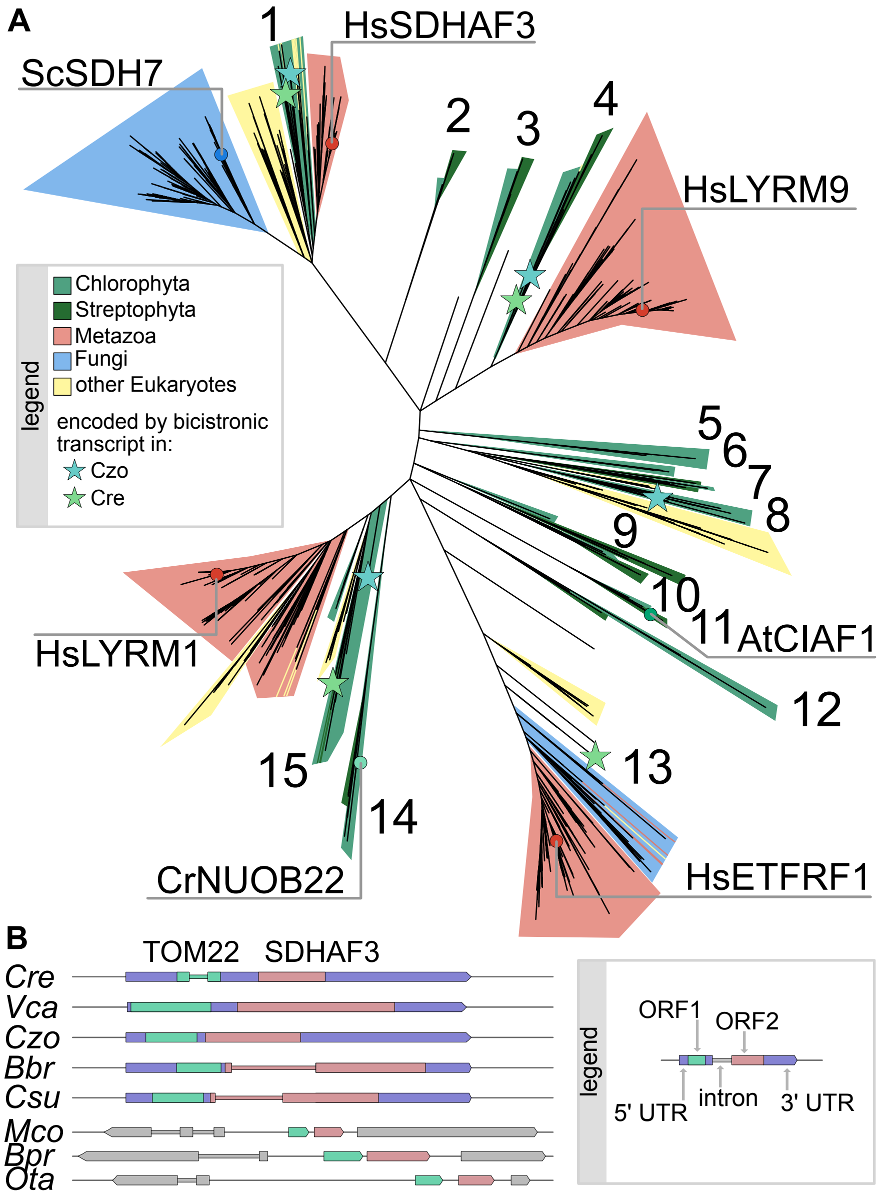


**Figure S8.** **Sequence analysis of the LYR-motif family.**

(**A**) Phylogenetic tree of LYR-motif family members from *C. reinhardtii* and *C. zofingiensis* and similar sequences from other species. Individual orthologous groups containing green algal sequences are numbered 1 – 15. Characterized homologs are labeled with protein names preceded by "Sc" for *Saccharomyces cerevisiae*, "Hs" for *Homo sapiens*, "Cr" for *C. reinhardtii*, and "At" for *Arabidopsis thaliana*. (**B**) Schematic of putative bicistronic loci from eight species. Each locus encodes a group 1 SDHAF3-like protein from group 1 in panel A. Purple bars indicate support from Iso-Seq or EST data for polycistronic transcription of both ORFs from a single transcript. Grey genes represent predicted unrelated gene models flanking the TOM22-like and SDHAF3-like ORFs. The species are as follows: *Cre = C. reinhardtii, Vca = V. carteri, Czo = C. zofingiensis, Bbr = Botryococcus braunii, Csu = Coccomyxa subellipsoidea, Mco = Micromonas commoda, Bpr = Bathycoccus prasinos, Ota = Ostreococcus tauri.*

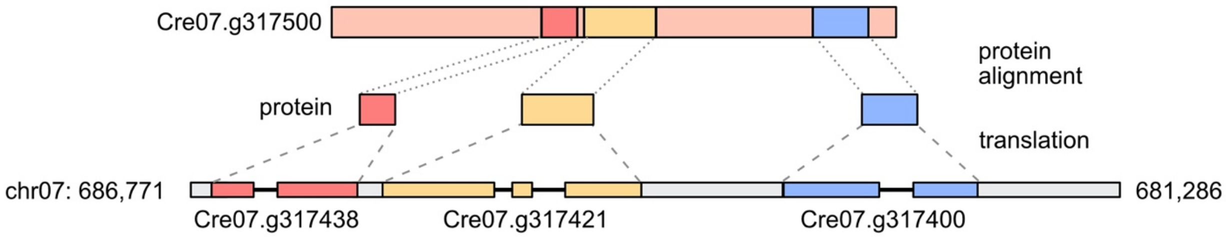


**Figure S9. An example of a putative duplication and fragmentation that results in a polycistronic locus.**

A subset of the polycistronic loci may be the result of a duplication of a monocistronically expressed gene (i.e., a transcript with one ORF), followed by fragmentation of the single ORF into two or more ORFs. Presented here is an example from *C. reinhardtii,* in which a duplication and fragmentation of Cre07.g317500 could have resulted in a three-ORF polycistronic locus. Regions of shared homology are color coded. Other examples of this duplication and fragmentation are described in Dataset S5.

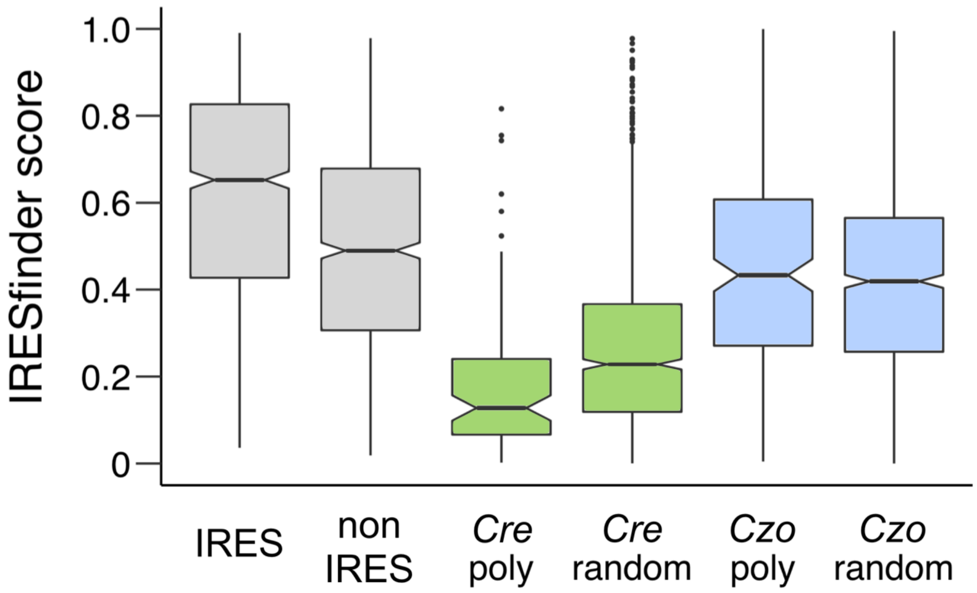


**Figure S10. Computational prediction of IRES function.**

In order to systematically evaluate whether polycistronic inter-ORF sequences function as IRESs (i.e. facilitate cap-independent translation of downstream ORFs), the inter-ORF sequences of all *C. reinhardtii* and *C. zofingiensis* polycistronic loci were subjected to analysis by a computational IRES prediction tool called IRESfinder. IRESfinder reports scores on a scale from 0 to 1 of indicating increasing likelihood that the sequences will function as an IRES. The distribution of these scores is presented here as a box plot for each set of sequences. See supplemental methods for details. As controls, sequences that had been empirically determined to function as IRESs ("IRES", *N* = 1000) or to not function as IRESs ("non IRES", *N* = 1000) were analyzed and their scores are plotted. All inter-ORF sequences from the polycistronic loci of *C. reinhardtii* (*N* = 87) and *C. zofingiensis* (*N* = 204) are plotted as "*Cre* poly" and "*Czo* poly", respectively. As an additional control, over 1000 sized matched sequences of random intergenic sequence from *C. reinhardtii* (*N* = 1044) and *C. zofingiensis* (*N* = 1020) are plotted as "*Cre* random" and "*Czo* random", respectively. ANOVA was performed to determine which differences are statistically significant. All sets were significantly smaller than the IRES set (*p* <0.01). *Cre* poly and *Czo* poly were not significantly different from their respective random counterparts. For each box plot, whiskers indicate 1.5 times the interquartile range, and notches indicate the confidence interval of the median. Outliers are presented as individual points.

## Supplemental Dataset Legends

**Dataset S1.** Master Table of Polycistronic loci
This Excel file contains detailed information on polycistronic loci identified in *C. reinhardtii* and *C. zofingiensis* including gene ID, coordinates, ORF size, inter-ORF spacing, and stop codon usage.

**Dataset S2.** Proteomics Analysis
Total protein from *C. reinhardtii* and *C. zofingiensis* we subjected to trypsin digest followed by mass spectrometry to identify expressed proteins. The results of this analysis are presented in this Excel file with separate tabs for each species. The columns are as follows:
"Peptide" => This column includes the amino acid sequence of a peptide, including both amino acids flanking the termini of the peptide. If multiple peptide sequences are consistent with the observed mass, these are listed in a coma-delimited list. Periods (".") indicate the termini of the peptide.
"Clean_Sequence" => Same as "Peptide" above, but flanking amino acids are removed.
"Gene ID" => The observed peptide is encoded by at least one isoform of this gene.
"Observation_Count" => The total spectral counts of the observed peptide.
"Unambiguous" => Is "Y" if the observed peptide is only encoded by one gene. Is "N" if more than one gene could encode the peptide.

**Dataset S3.** *In Vitro* Transcription/Translation Constructs
This Excel file contains the sequences and related information for all constructs that were synthesized for use in the coupled *in vitro* transcription and translation experiments.

**Dataset S4.** Homology Analysis

The protein sequences of polycistronically-expressed proteins in *C. reinhardtii* and *C. zofingiensis* were used to identify potential homologous colinear gene pairs in five other chlorophyte species. This Excel file contains information on colinear gene pairs in the others species including gene IDs, coordinates, and support from EST or Iso-Seq data.

**Dataset S5.** Functional Annotations
This Excel file contains manually curated data on the predicted function of polycistronically expressed proteins.

**Dataset S6.** PredAlgo Analysis
This Excel file contains predictions about the sub-cellular targeting of all proteins in the v5.7 annotations of *C. reinhardtii* and the v5.3 annotations of *C. zofingiensis* as generated by PredAlgo using default parameters.

**Dataset S7**. IRESfinder Analysis
The polycistronic inter-ORF sequences of *C. reinhardtii* (Cre_poly) and *C. zofingiensis* (Czo_poly) were subjected to analysis by IRESfinder in mode 0 with default settings (See SI Methods for details). Over 1000 size-matched stretches of random intergenic sequence from *C. reinhardtii* (Cre_random) and *C. zofingiensis* (Czo_random) were analyzed as controls. Additionally, 1000 sequences determined to have IRES function (IRES) and 1000 sequences determined not to have IRES function (non_IRES) were analyzed. Each sequence and its IRESfinder score is presented in the corresponding tab.

**Dataset S8.** Annotations for *C. reinhardtii* (v5.7)
The *C. reinhardtii* structural gene annotations for that were used in this work in gff3 format.

**Dataset S9**. Annotations for *C. zofingiensis* (v5.3)

The *C. zofingiensis* structural gene annotations for that were used in this work in gff3 format.

## Supplemental References

1. S. V. Novoselov, *et al.*, Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii. *EMBO J.* **21**, 3681–3693 (2002).
2. M. Mariotti, A. V Lobanov, R. Guigo, V. N. Gladyshev, SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res.* **41**, e149 (2013).
3. M. Mariotti, A. V Lobanov, R. Guigo, V. N. Gladyshev, SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res.* **41**, e149 (2013).
4. M. Tardif, *et al.*, PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* **29**, 3625–3639 (2012).
5. J. Zhao, *et al.*, IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. *J. Genet. Genomics* **45**, 403–406 (2018).
6. R. J. Craig, A. R. Hasan, R. W. Ness, P. D. Keightley, Comparative genomics of Chlamydomonas. *bioRxiv*, 2020.06.13.149070 (2020).
7. F. R. Cross, Tying Down Loose Ends in the Chlamydomonas Genome: Functional Significance of Abundant Upstream Open Reading Frames. *G3 Genes, Genomes, Genet.* **6**, 435–446 (2016).
8. D. Strenkert, S. Schmollinger, F. Sommer, M. Schulz-Raffelt, M. Schroda, Transcription factor-dependent chromatin remodeling at heat shock and copper-responsive promoters in Chlamydomonas reinhardtii. *Plant Cell* **23**, 2285–301 (2011).
9. D. Strenkert, *et al.*, Multiomics resolution of molecular events during a day in the life of Chlamydomonas. *Proc. Natl. Acad. Sci.* **116**, 2374–2383 (2019).
10. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
11. Y. Zhang, *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9** (2008).
12. Y. Shen, Y. Liu, L. Liu, C. Liang, Q. Q. Li, Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in Chlamydomonas reinhardtii. *Genetics* **179**, 167–176 (2008).
13. C. D. Silflow, R. L. Chisholm, T. W. Conner, L. P. Ranum, The two alpha-tubulin genes of Chlamydomonas reinhardi code for slightly different proteins. *Mol. Cell. Biol.* **5**, 2389 LP – 2398 (1985).
14. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–70 (2011).

15. S. S. Merchant, *et al.*, The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science (80-. ).* **318**, 245–250 (2007).

16. M. S. Roth, *et al.*, Chromosome-level genome assembly and transcriptome of the green alga Chromochloris zofingiensis illuminates astaxanthin production. *Proc. Natl. Acad. Sci. U. S. A.* **114** (2017).

17. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

18. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

19. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

20. I. K. Blaby, *et al.*, Systems-level analysis of nitrogen starvation-induced modifications of carbon metabolism in a Chlamydomonas reinhardtii starchless mutant. *Plant Cell* **25**, 4305–23 (2013).

21. D. Malasarn, *et al.*, Zinc deficiency impacts CO2 assimilation and disrupts copper homeostasis in Chlamydomonas reinhardtii. *J Biol Chem* **288**, 10672–10683 (2013).

22. J. Kropat, *et al.*, Copper economy in Chlamydomonas : Prioritized allocation and reallocation of copper to respiration vs. photosynthesis. *Proc. Natl. Acad. Sci.* **112**, 2644–2651 (2015).

23. M. Castruita, *et al.*, Systems biology approach in Chlamydomonas reveals connections between copper nutrition and multiple metabolic steps. *Plant Cell* **23**, 1273–1292 (2011).

24. E. I. Urzica, *et al.*, Remodeling of Membrane Lipids in Iron-starved Chlamydomonas. *J. Biol. Chem.* **288**, 30246–30258 (2013).

25. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

26. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

27. C. Trapnell, *et al.*, Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).

28. S. D. Gallaher, *et al.*, High-throughput sequencing of the chloroplast and mitochondrion of Chlamydomonas reinhardtii to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. *Plant J.* **93** (2018).

29. J. A. Nelson, P. B. Savereide, P. A. Lefebvre, The CRY1 gene in Chlamydomonas reinhardtii: structure and use as a dominant selectable marker for nuclear transformation. *Mol. Cell. Biol.* **14**, 4011–4019 (1994).

30. A. Rekas, J. R. Alattia, T. Nagai, A. Miyawaki, M. Ikura, Crystal structure of venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity. *J. Biol. Chem.* **277**, 50573–50578 (2002).

31. L.-X. Shi, S. M. Theg, Energetic cost of protein import across the envelope membranes of chloroplasts. *Proc. Natl. Acad. Sci.* **110**, 930–935 (2013).

32. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

33. S. Weingarten-Gabbay, *et al.*, Comparative genetics: Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-. ).* **351** (2016).

34. J. Wang, M. Gribskov, IRESpy: An XGBoost model for prediction of internal ribosome entry sites. *BMC Bioinformatics* **20**, 409 (2019).

35. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).