doi:10.1111/tpj.13788

Supplemental Figures 1 – 8

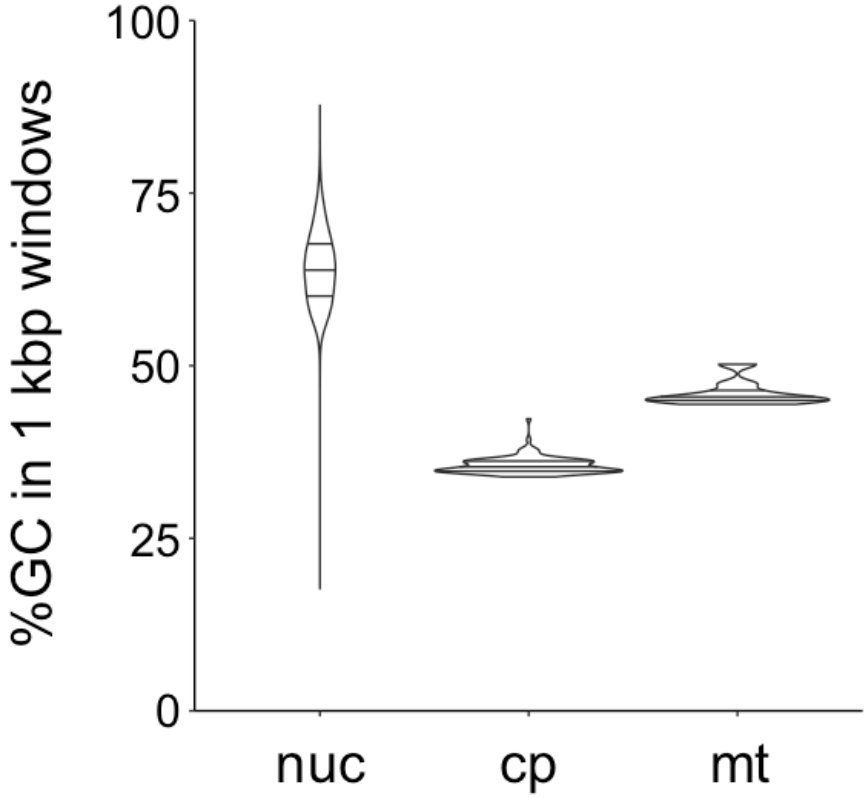


Figure S1 – GC Content

The distribution of GC content in 1 kbp windows for the nuclear, chloroplast and mitochondrial genomes is presented as a violin plot. Horizontal bars indicate the 75, 50, and 25% quartiles.

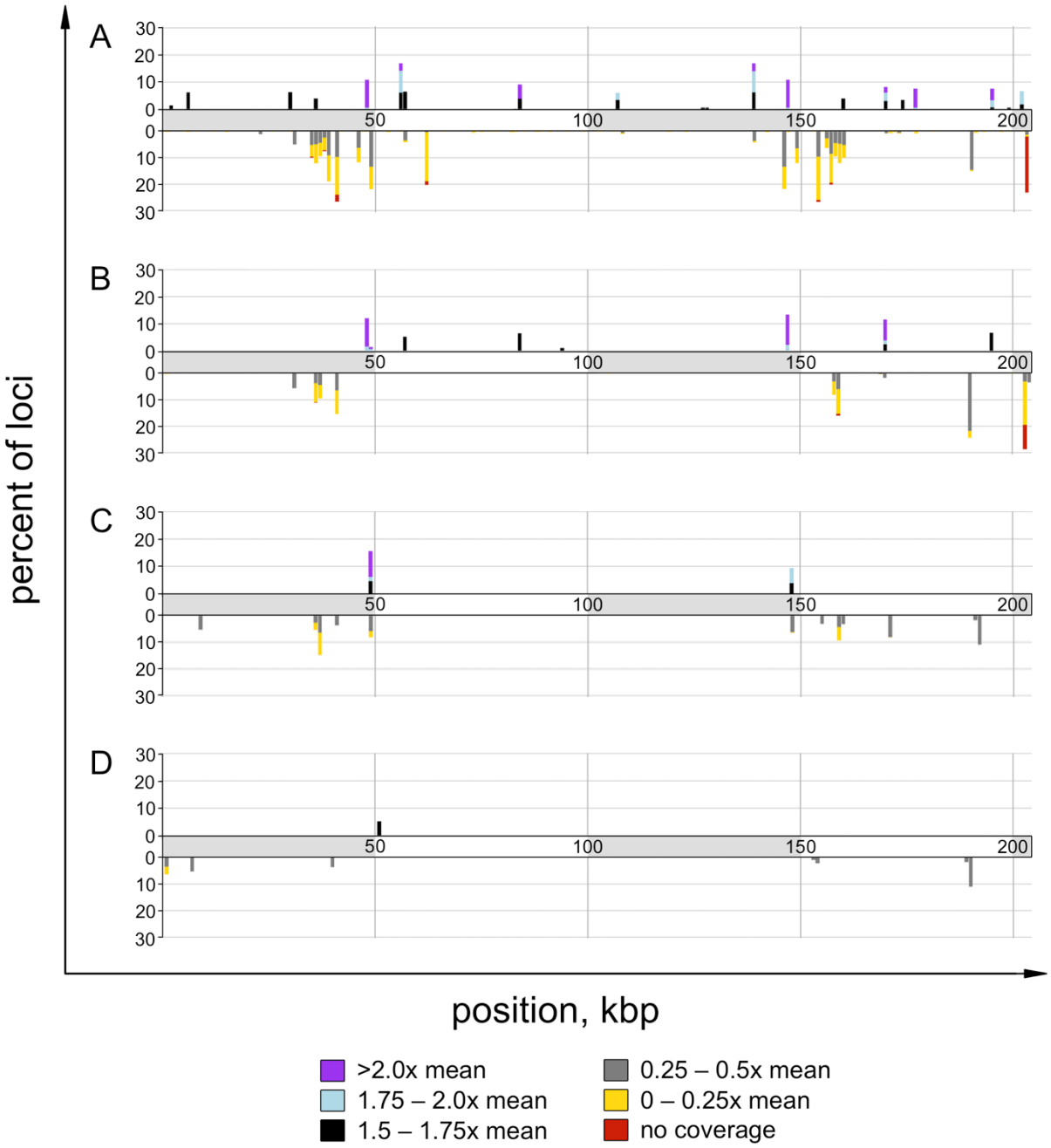


Figure S2 – Heterogeneity of coverage depth in 1 kbp windows

A set of $1x10^8$ paired-end DNA-Seq reads (100 +100 nt) were aligned to each of four genome versions in parallel using the same parameters. The mean depth of coverage across each genome version was determined. Then, for each locus within non-overlapping 1 kbp windows, the percentage of loci with coverage depths significantly

different than the mean were calculated and plotted. The percentage of loci greater than 1.5 x the mean is plotted above line, and the percentage of loci less than 0.5 x the mean is plotted below. Each bar is further sub-divided and color coded as indicated by the legend. The four genome versions are as follows: (**A**) BK000554.2 from NCBI GenBank 2002, (**B**) FJ423446.1 from NCBI Genbank 2009, (**C**) cv11, which is a variant-corrected reconstruction of FJ423446.1, and (**D**) CPv4, which is a *de novo* assembly.

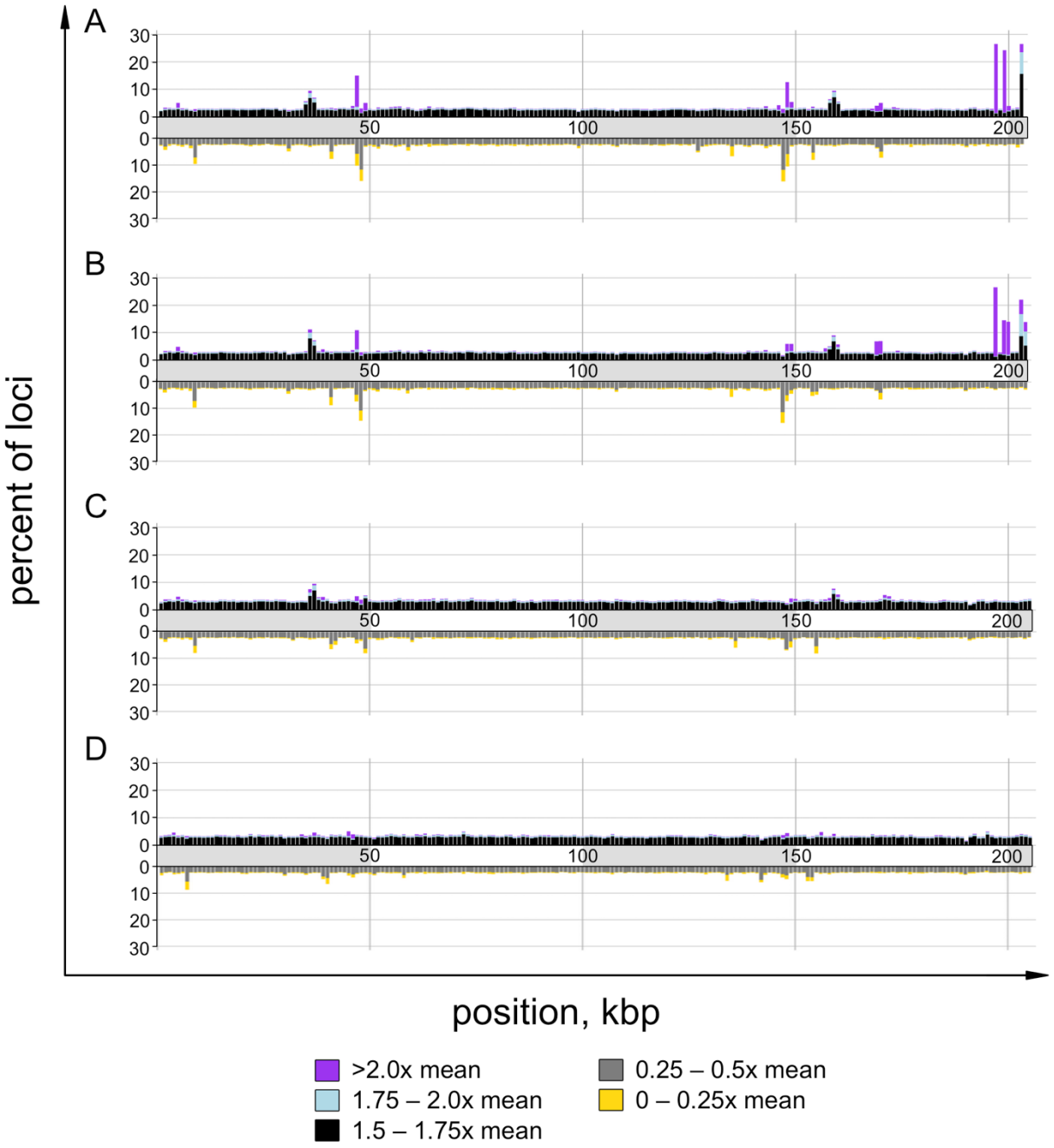


Figure S3 – Heterogeneity of inferred DNA-Seq fragment sizes

A set of $1x10^8$ paired-end DNA-Seq reads (100 +100 nt) were aligned to each of four genome versions in parallel using the same parameters. The fragment size was inferred by the relative alignment of the two sequencing reads for each fragment of DNA. Next, a mean was determined for the full set of aligned reads, and individually for each locus. For each locus within non-overlapping 1 kbp windows, the percentage of loci with

inferred fragment sizes significantly different than the mean were calculated and plotted. The percentage of loci greater than 1.5 x the mean is plotted above line, and the percentage of loci less than 0.5 x the mean is plotted below. Each bar is further sub-divided and color coded as indicated by the legend. The four genome versions are as follows: (**A**) BK000554.2 from NCBI GenBank 2002, (**B**) FJ423446.1 from NCBI GenBank 2009, (**C**) cv11, which is a variant-corrected reconstruction of FJ423446.1, and (**D**) CPv4, which is a *de novo* assembly.

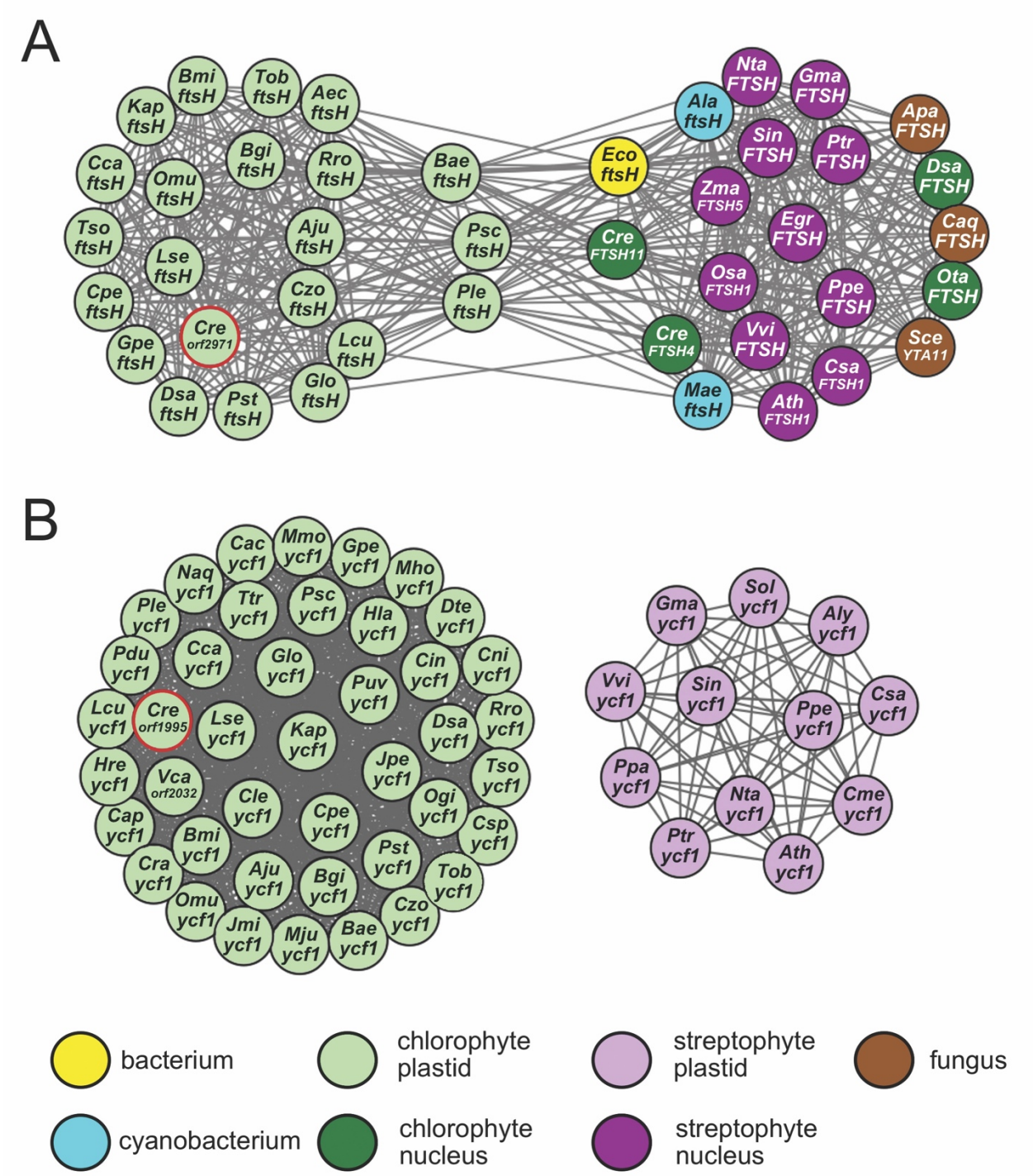


Figure S4 – Protein similarity networks inform annotation of ORFs

A protein similarity network was constructed from a BLASTP analysis (all versus all, E-value cutoff = $1x10^{-10}$) for two of the previously unannotated *C. reinhardtii* chloroplast ORFs compared with annotated genes in other species. Each node represents a protein

encoded in either the nucleus or plastid of various species, and is labeled with a three letter code for the species and the assigned gene name. Details and accession numbers are given in Supplemental Table S02. Nodes are color coded as indicated by the legend. The two unannotated *C. reinhardtii* chloroplast ORFs are highlighted in red. (**A**) Comparison of *C. reinhardtii orf2971* to *ftsH* orthologs and related genes in 47 species. (**B**) Comparison of *C. reinhardtii orf1995* with *ycf1* orthologs in 59 other species.

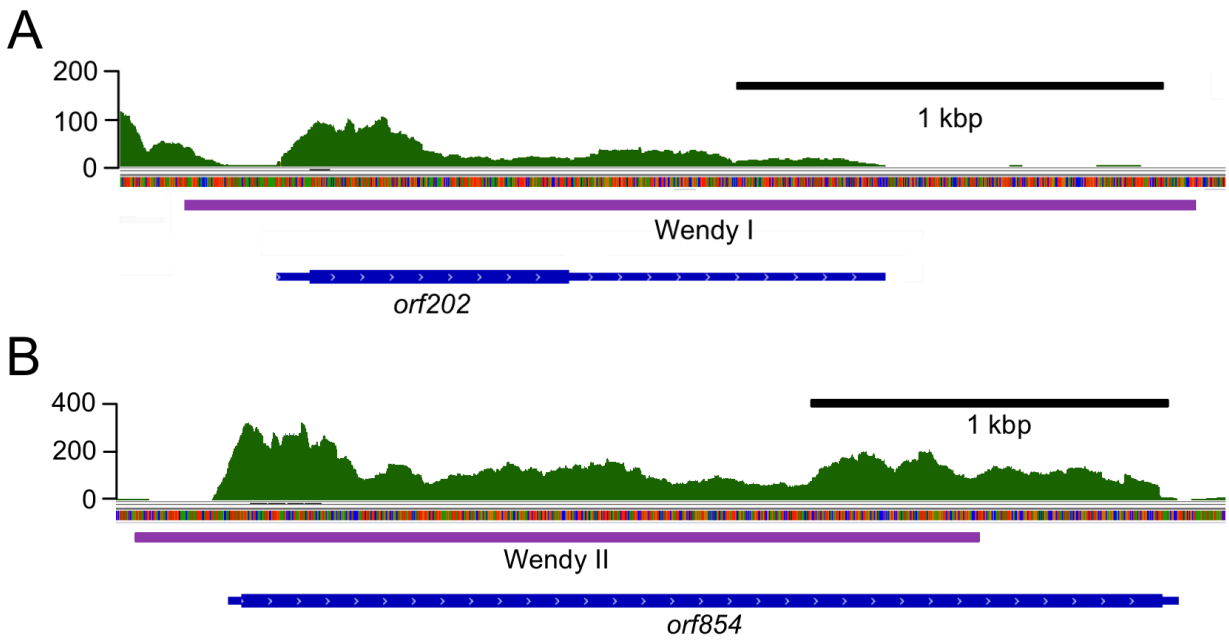
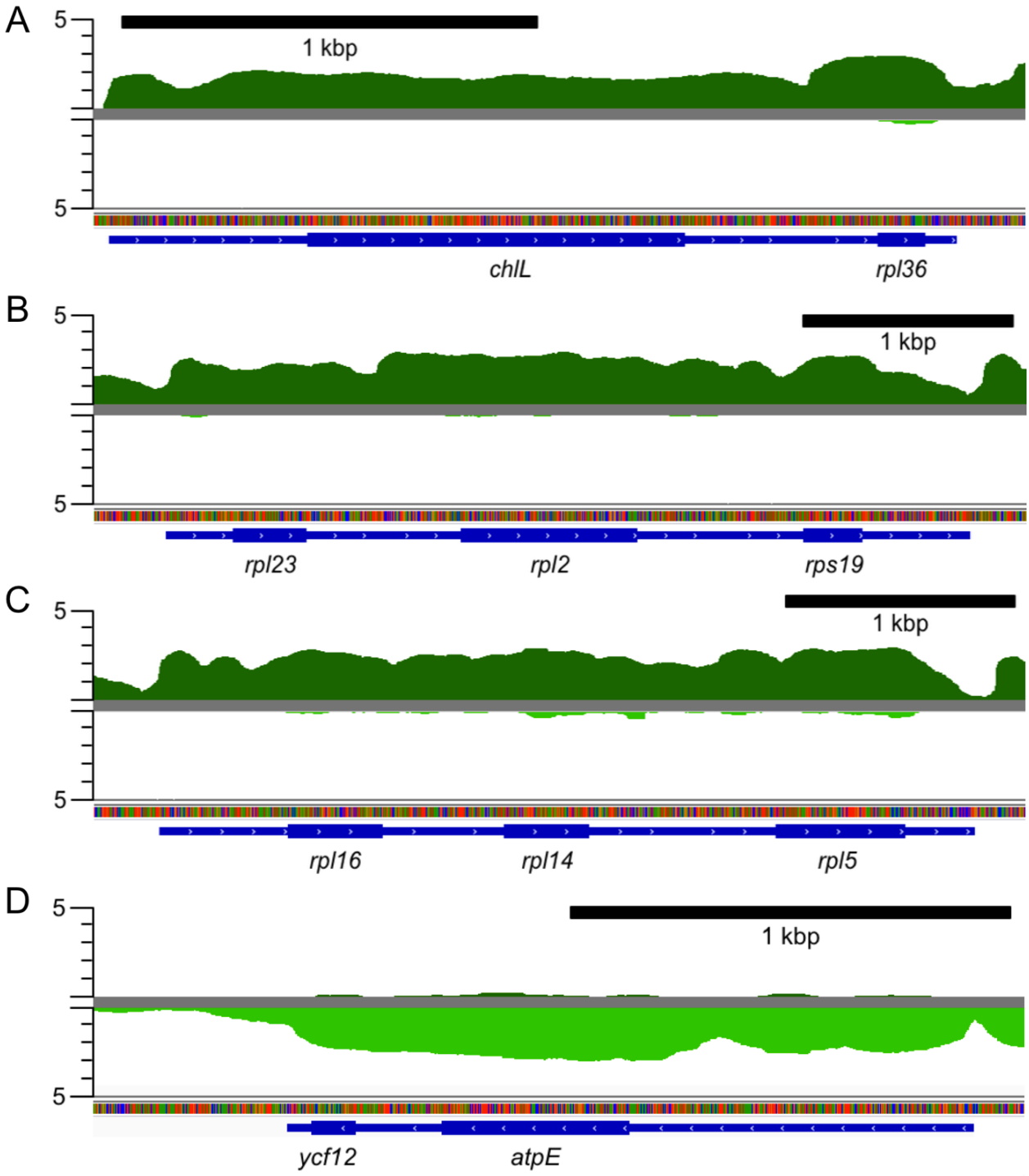
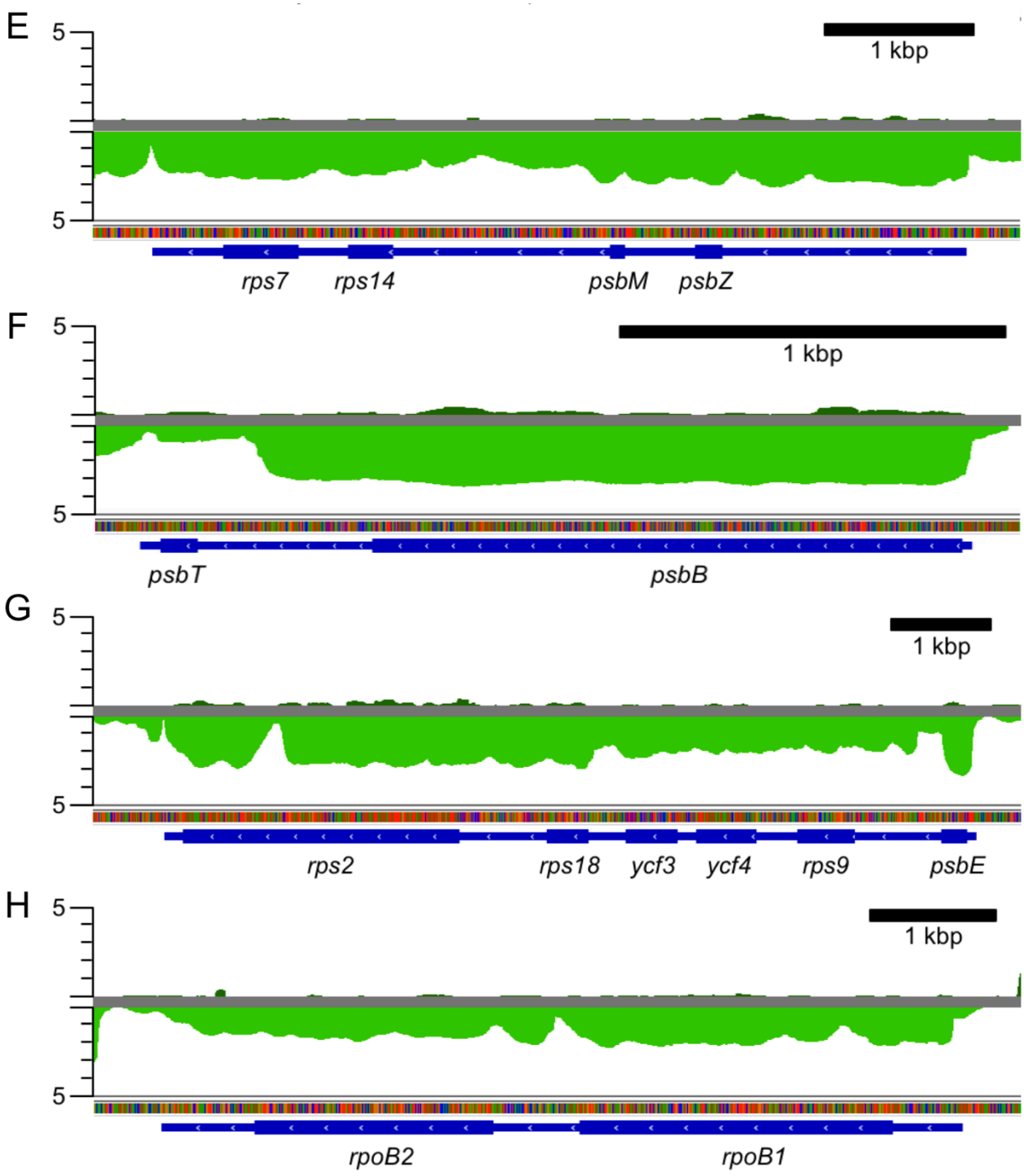
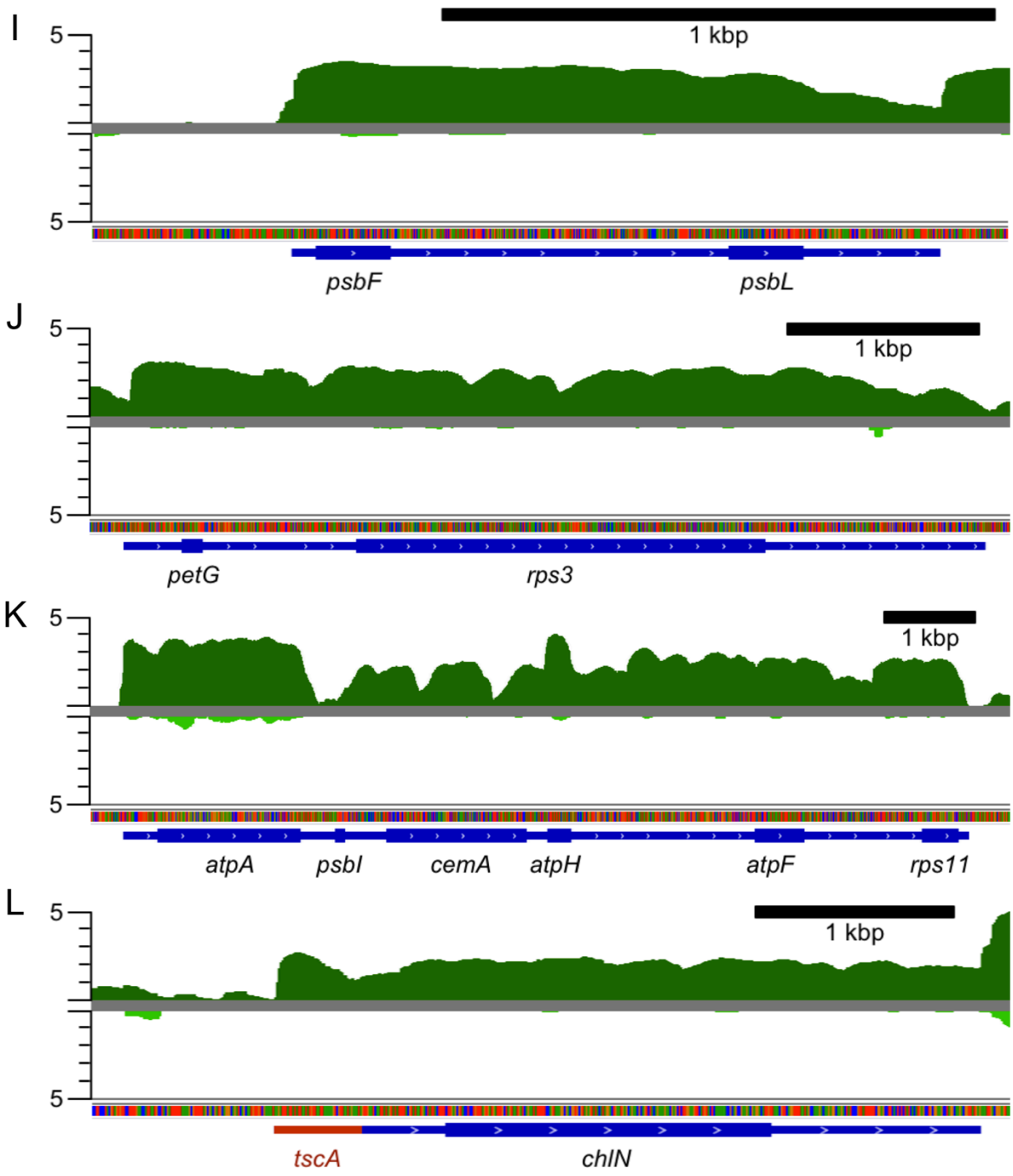


Figure S5 Evidence for expression from ORFs in Wendy Transposons

The *C. reinhardtii* chloroplast carries two copies of the Wendy transposon. (**A**) Wendy I has a ORF that is predicted to encode a 202 aa polypeptide. (**B**) Wendy II is predicted to encode an 854 aa polypeptide. The Wendy II ORF extends beyond the boundary of the transposon sequence. In both panels, the transposon sequence is represented by a purple bar. The putative transcript is shown by a blue bar, with a thicker portion to indicate the ORF. RNA-Seq coverage from a representative experiment is shown in green on a linear scale above the sequence. A y-axis on the right-hand side indicates depth of coverage. The scale bar indicates 1 kbp of sequence.


A
5
1 kbp
5
chlL
rpl36
B
5
1 kbp
5
rpl23
rpl2
rps19
C
5
1 kbp
5
rpl16
rpl14
rpl5
D
5
1 kbp
5
ycf12
atpE


E
5
5
1 kbp
rps7
rps14
psbM
psbZ
F
5
5
1 kbp
psbT
psbB
G
5
5
1 kbp
rps2
rps18
ycf3
ycf4
rps9
psbE
H
5
5
1 kbp
rpoB2
rpoB1


I
5
1 kbp
5
psbF
psbL
J
5
1 kbp
5
petG
rps3
K
5
1 kbp
5
atpA
psbI
cemA
atpH
atpF
rps11
L
5
1 kbp
5
tscA
chlN

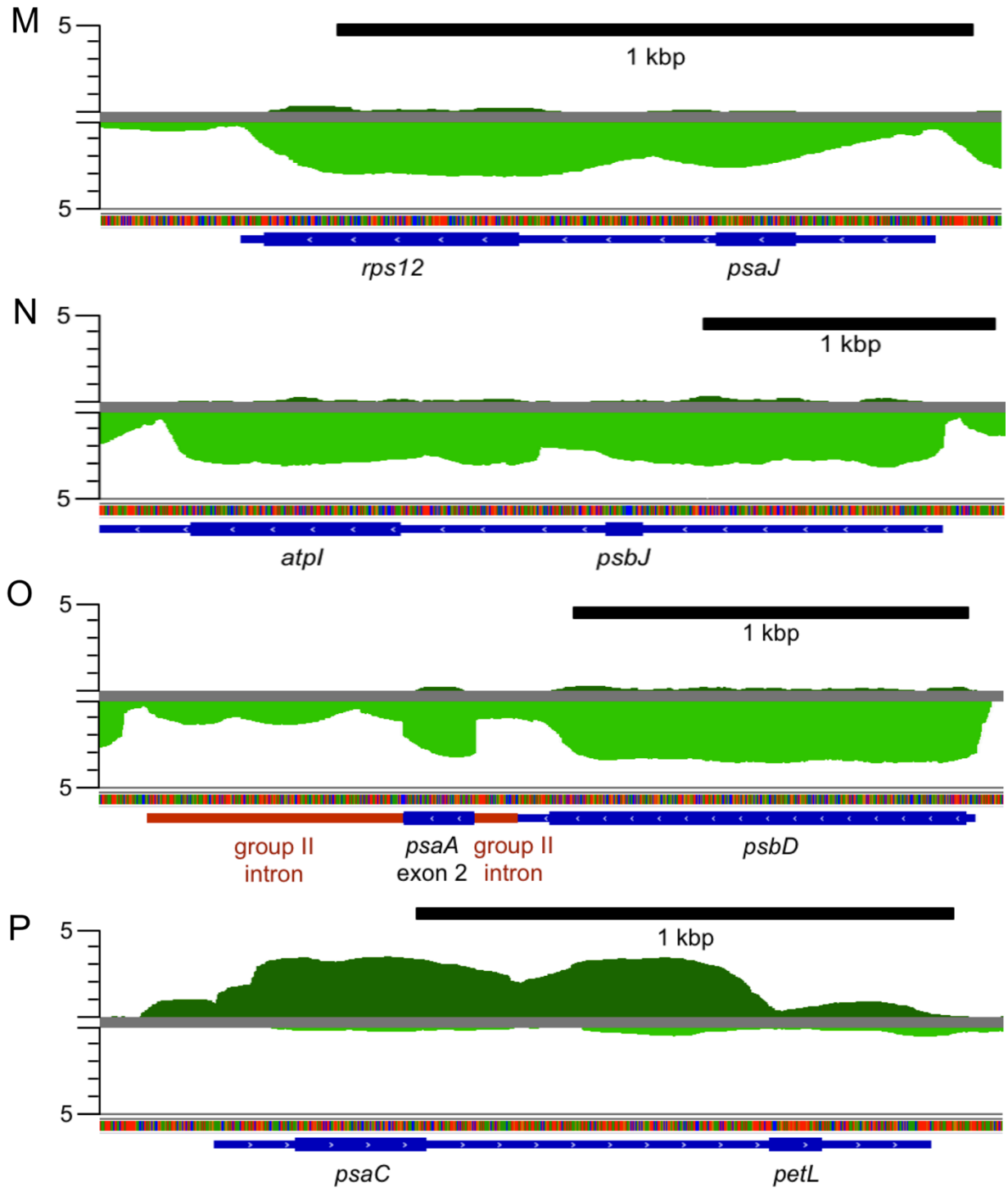

M
5
1 kbp
5
rps12
psaJ
N
5
1 kbp
5
atpI
psbJ
O
5
1 kbp
5
group II intron
psaA exon 2
group II intron
psbD
P
5
1 kbp
5
psaC
petL

Figure S6 – Evidence for polycistronic expression of *C. reinhardtii* chloroplast genes

Strand-specific RNA-Seq coverage from a representative RNA-Seq experiment is presented for clusters of genes. Log_{10}-transformed coverage from the plus strand is presented above the line in dark green, and minus strand coverage is presented below the line in light green. A scale bar to the right indicates five orders of magnitude. Protein coding genes are presented in blue, with thicker portions to indicate CDSs and thinner portions to indicate UTRs. The direction of transcription is indicated by white arrowheads. Introns, when present, are indicated in red. A black scale bar in each panel indicates 1 kbp. Gene clusters are as follows: (**A**) *chl – rpl36*, (**B**) *rpl23 – rpl2 – rps19*, (**C**) *rpl 16 – rpl 14 – rpl5*, (**D**) *atpE – ycf12*, (**E**) *psbZ – psbM – rps14 – rps7*, (**F**) *psbB – psbT*, (**G**) *psbE – rps9 – ycf4 – ycf3 – rps18 – rps2*, (**H**) *rpoB1 – rpoB2* (**I**) *psbF – psbL*, (**J**) *petG – rps3*, (**K**) *atpA – psbI – cemA – atpH – atpF – rps11*, (**L**) *tscA – chlN*, (**M**) *psaJ – rps12*, (**N**) *psbJ – atpI*, (**O**) *psbD – psaA* exon2, (**P**) *psaC – petL*.

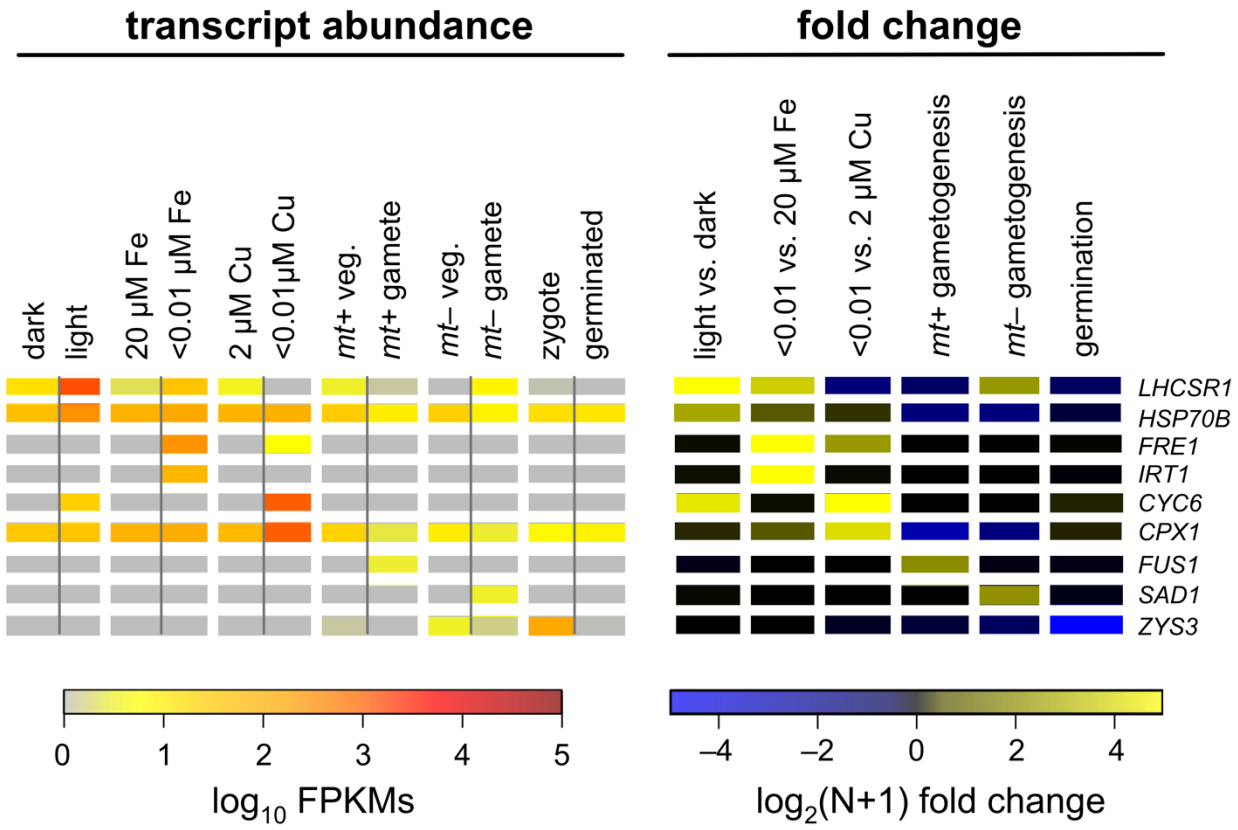
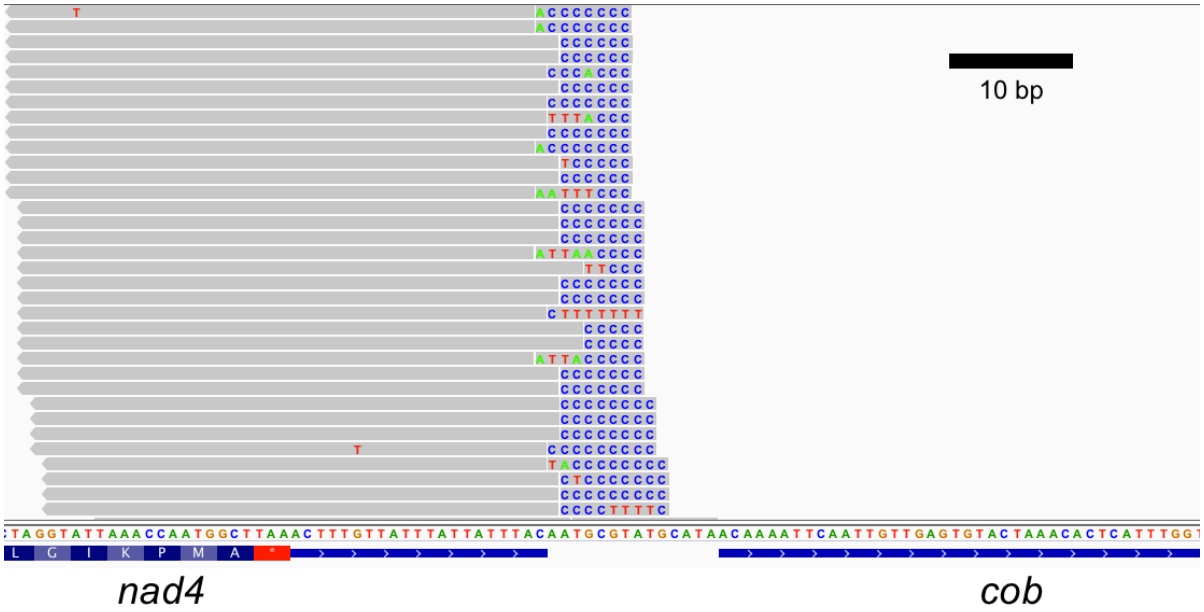


Figure S7 – Expression of sentinel genes validate the RNA-Seq Analysis

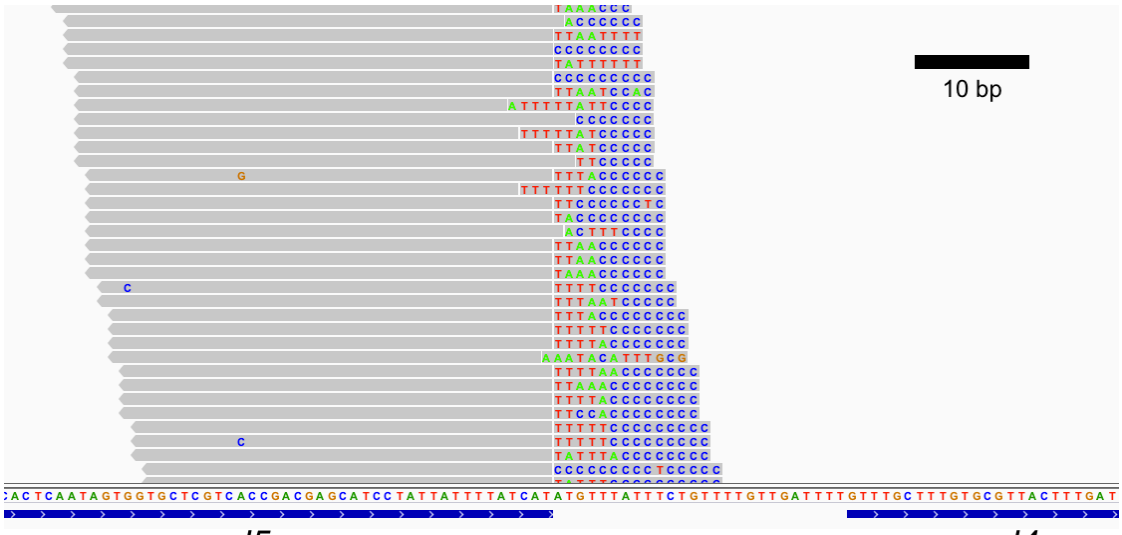
Selected sentinel genes encoded in the nucleus were analyzed in parallel with the study presented in main Figures 7 and 8. Transcript abundance determinations in terms of $\log_{10}$-transformed FPKMs are presented in the heatmap on the left. The $\log_2$(N+1) fold change between pairs of samples is presented in the heatmap on the right. Sentinel genes are as follows: *LHCSR1* and *HSP70B* are expressed upon transition from dark to light. *FRE1* and *IRT1* are expressed in response to low Fe. *CYC6* and *CPX1* are expressed in response to low Cu. *FUS1* is expressed in *mt*+ gametes. *SAD1* is expressed in *mt*– gametes. *ZYS3* is expressed in zygotes.

A

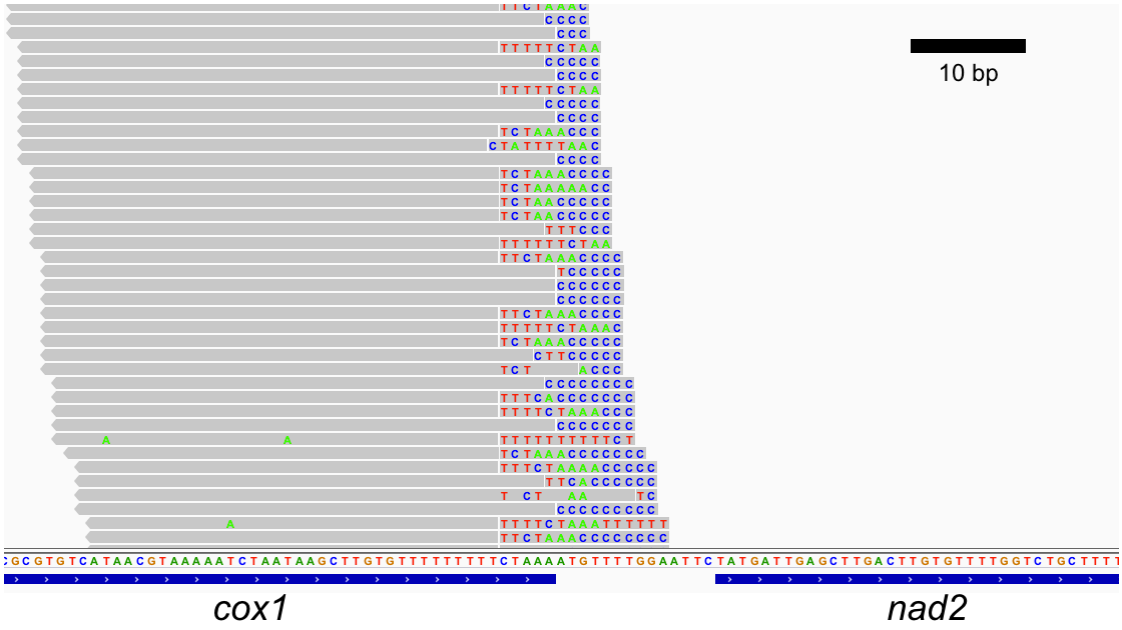
B



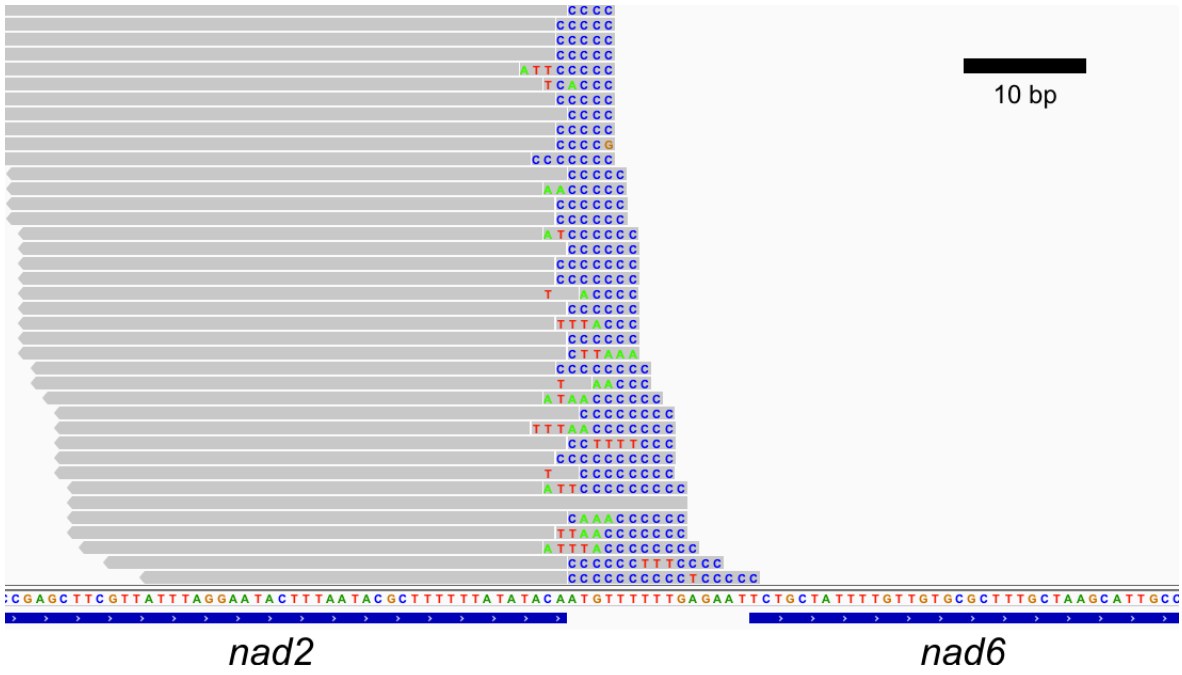
C



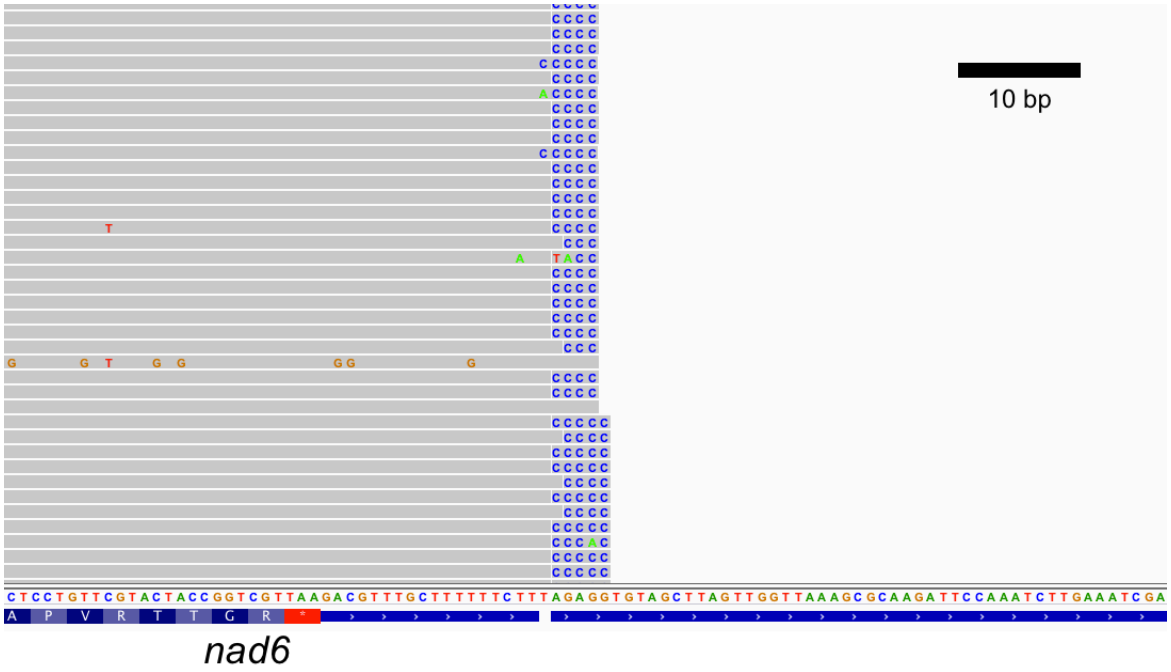
D



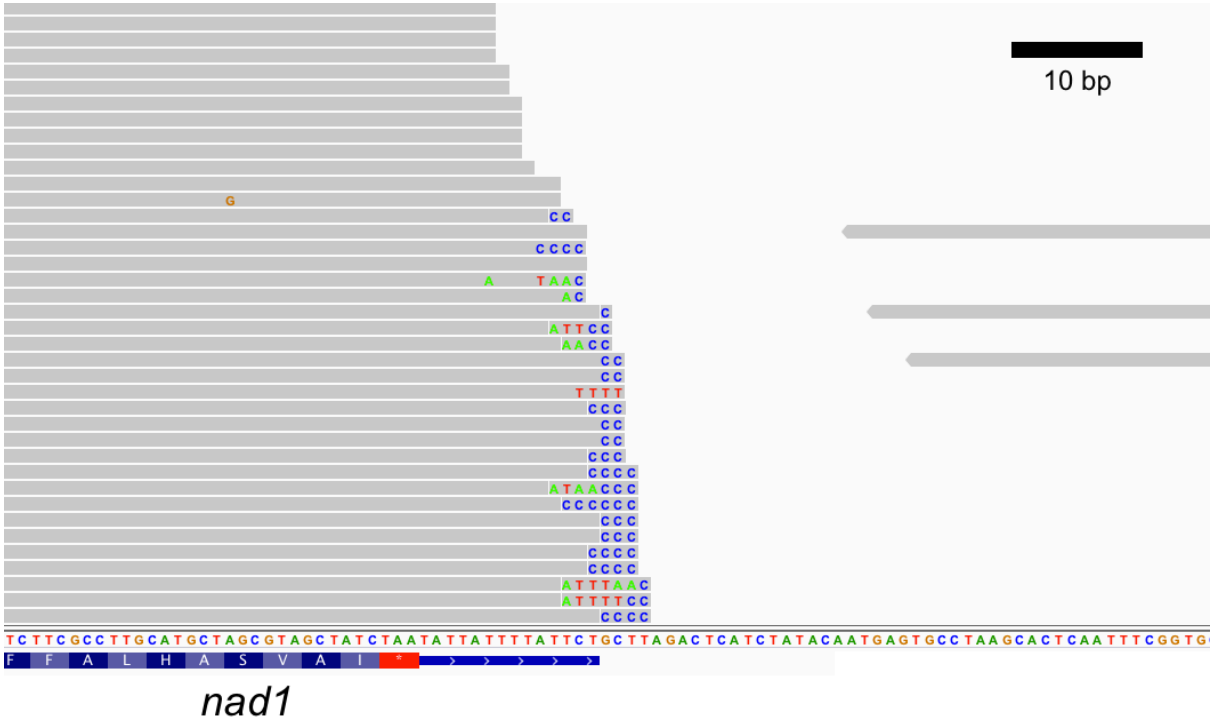
E



F



G



H

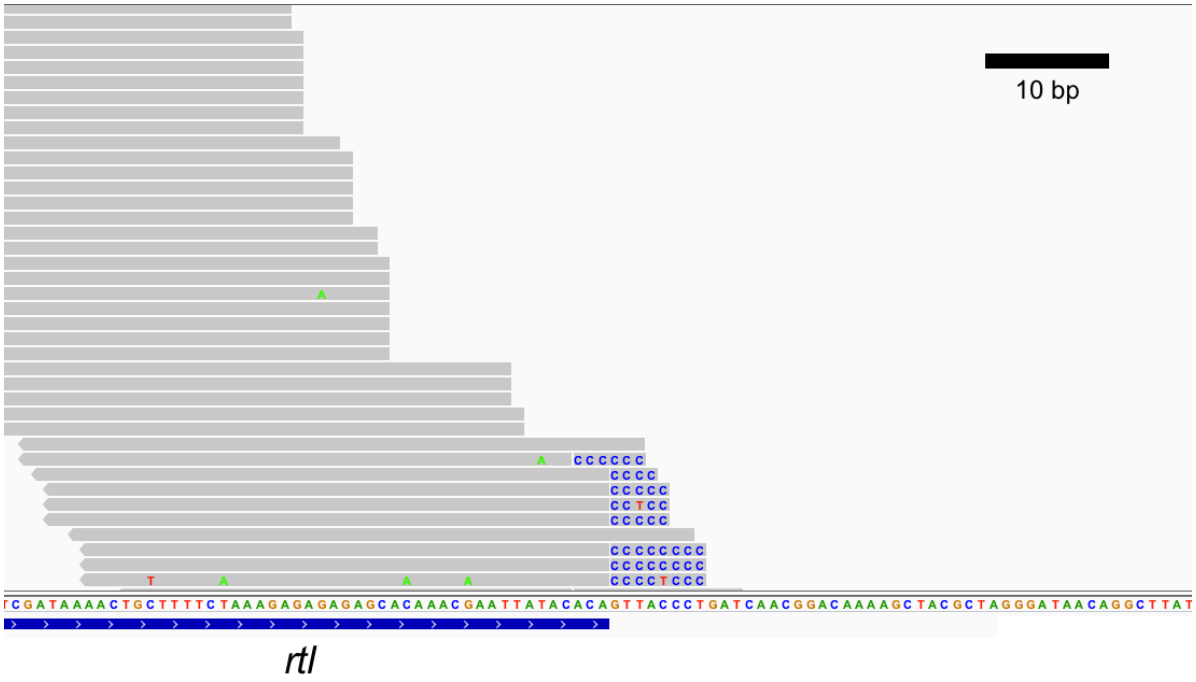


Figure S8 – Evidence for polynucleotide 3’ tail on mitochondrial transcripts

The alignment of RNA-Seq reads to all eight mitochondrial protein-encoding genes suggest heterogeneous polynucleotide tails are common at the 3’ ends of the transcripts. Shown are 50 nt RNA-Seq reads aligned to MTv4 with RNA-STAR and visualized with IGV. The reference sequence is presented at the bottom of each panel. Base calls in the RNA-Seq reads that are in agreement with the reference are gray. Mismatches are indicated as follows: green = A, blue = C, orange = G, and red = T. Each panel is arranged so that the gene is encoded on the plus strand. Shown are the 3’ ends of (**A**) *cob*, (**B**) *nad4*, (**C**) *nad5*, (**D**) *cox1*, (**E**) *nad2*, (**F**) *nad6*, (**G**) *nad1*, and (**H**) *rtl*. A size bar in each panel indicates 10 bp.