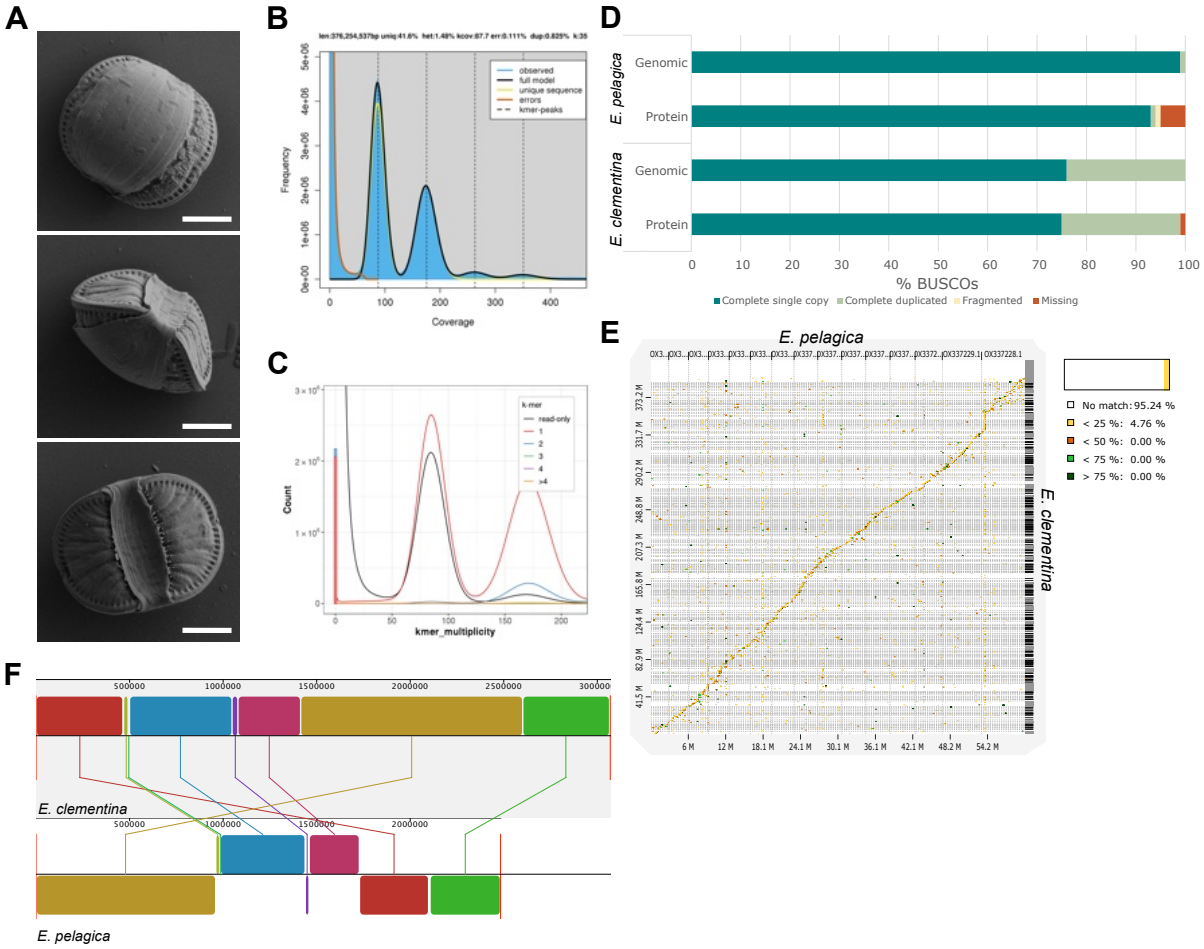


Figure S1. *E. clementina* genome assembly statistics and features
(A) Scanning electron micrographs of *E. clementina*, scale bar 5µm. Top: View looking down on the dorsal girdle band. Middle: View down the apical axis. Bottom: View of the ventral face, lined by prominent fenestral bars regularly spaced between the radial striae. The raphe lies along the strongly curved keel on the ventral margin and pinches slightly towards the dorsal margin.
(B) GenomeScope spectrum of 35-mer multiplicity collected from the Illumina sequencing reads. Peak at 1x coverage (~90) and 2x coverage (~180), consistent with a diploid genome.
(C) Merqury spectrum of k-mer multiplicity collected from the Illumina sequencing reads, stacked lines colored by number of times k-mer is seen in the genome assembly. Few k-mers within the heterozygous and homozygous peaks are read-only (black), suggesting that the assembly is not missing significant sequence present in the reads.
(D) Stramenopile-specific Benchmarking Universal Single-Copy Orthologs (BUSCOs) for *E. pelagica* and *E. clementina* genomes and proteomes. Both genomes contain all stramenopile BUSCOs, however the *E. pelagica* annotation is less complete. The genome and proteome of *E. clementina* show some duplication.
(E) Whole genome alignment of the *E. clementina* and *E. pelagica* genome assemblies. White indicates no sequence homology, yellow indicates alignments at <25% nucleotide identity. There is only 4.76% sequence homology between the two genomes at the nucleotide level, all at <25% identity.
(F) Genomic synteny between the whole genome alignments of the *E. clementina* and *E. pelagica* diazoplasts, showing 7 syntenic blocks.

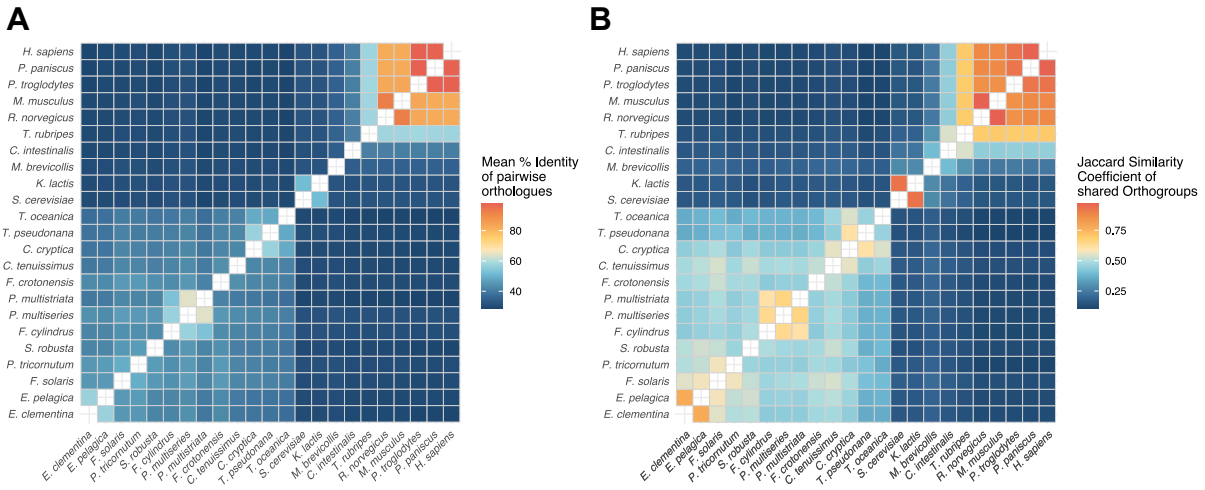


Figure S2. Detailed gene family divergence statistics

(A) Heat map showing mean percent amino acid identity of pairwise orthologs between all species used for comparative analysis.

(B) Same as A, showing the Jaccard similarity coefficient of the shared orthogroup overlap.

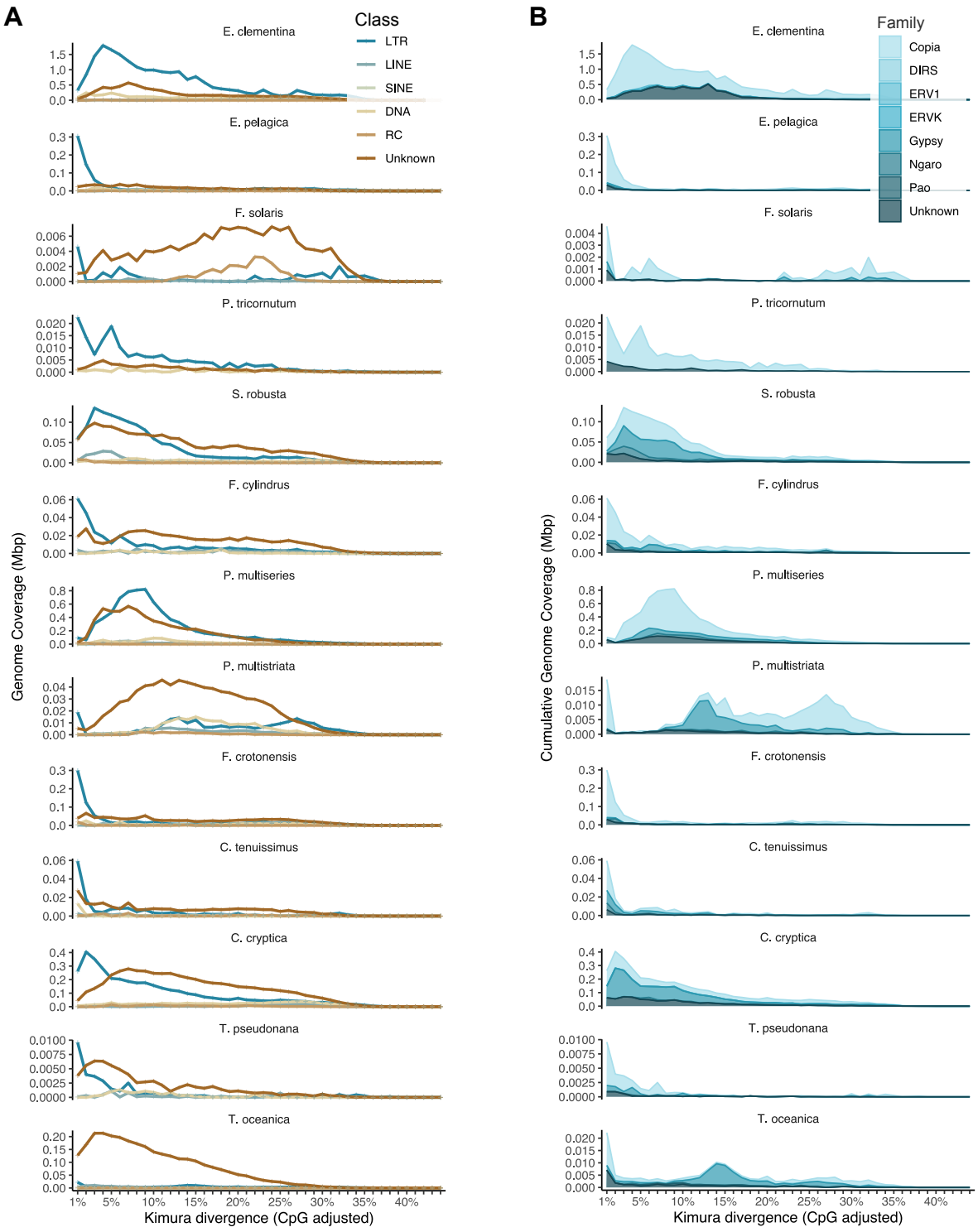


Figure S3. Repeat Landscapes across all diatoms

(A) Repeat landscape plots for all diatoms used for comparative analysis. Amount of the genome occupied by repeats plotted by divergence from inferred ancestral sequence. More divergence suggests an older insertion. Genome coverage is plotted on a free-y axis scale to display the full repeat expansion dynamics for each diatom.

(B) Stacked repeat landscape plots for LTR elements, colored by family.

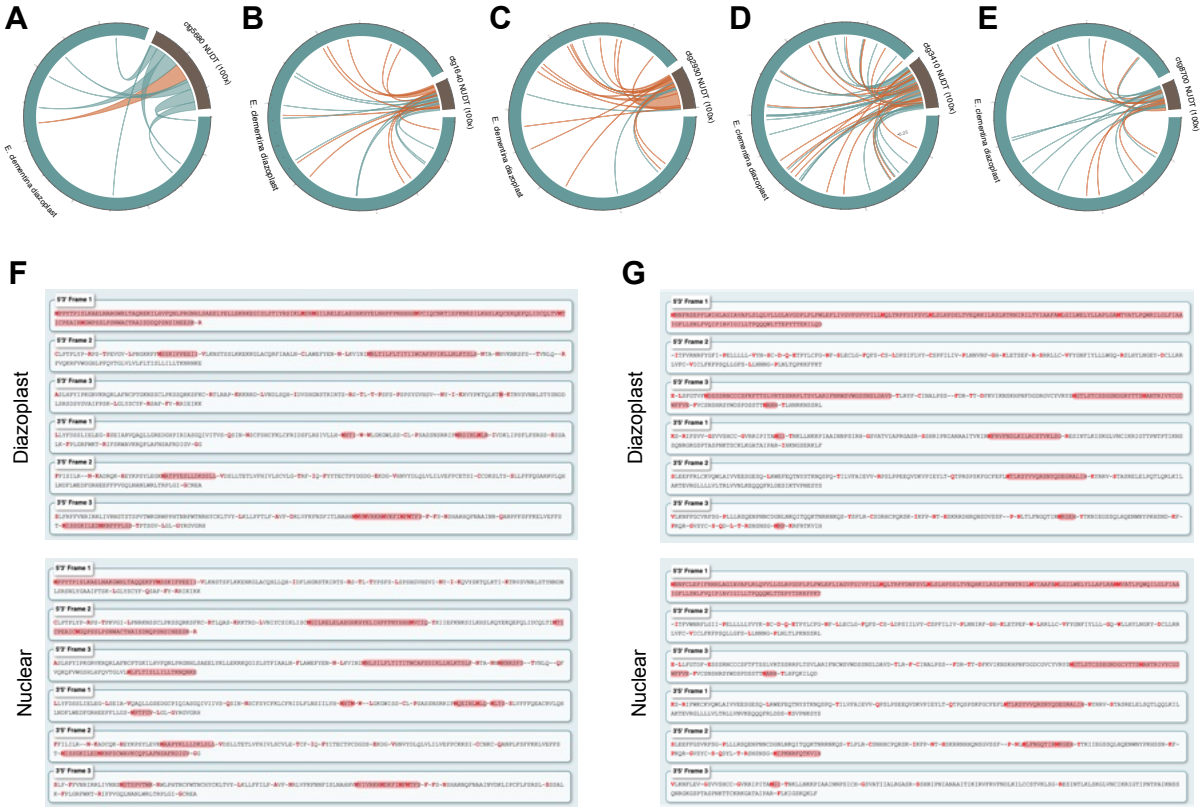


Figure S4. NUDT fragmentation and gene containing regions

(A-E) Circlize plots depicting the fragmentation and rearrangement of the NUDTs. The diazoplast genome (blue) and the NUDT on labeled contig (brown) with chords connecting source diazoplast regions to their corresponding nuclear region, inversions in red. The length of the NUDT is depicted at 100x true relative length for ease of visualization.

(F) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (transcriptional repressor, gene ID: P3f56_RS08570). The copy within the NUDT (bottom) is untruncated (100% of the full-length gene) by nucleotide sequence and is 96% identical to the corresponding diazoplast gene (top). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has a mutation that results in a premature stop codon at amino acid 39 (out of 177). Red highlight indicates a potential translation. 5'3' Frame 1 is the native diazoplast frame.

(G) Translation in all potential frames of the gene contained within the NUDT on contig ctg002090 (low-complexity tail membrane protein, gene ID: P3F56_RS01750). The gene is 9% truncated at the 3' terminus (91% of the full-length gene). Compared to the diazoplast-encoded gene, the gene contained in the NUDT has several non-synonymous mutations and is missing 16 amino acids at the C-terminus. Red highlight indicates a potential translation. 5'3' Frame 1 is the native diazoplast frame.

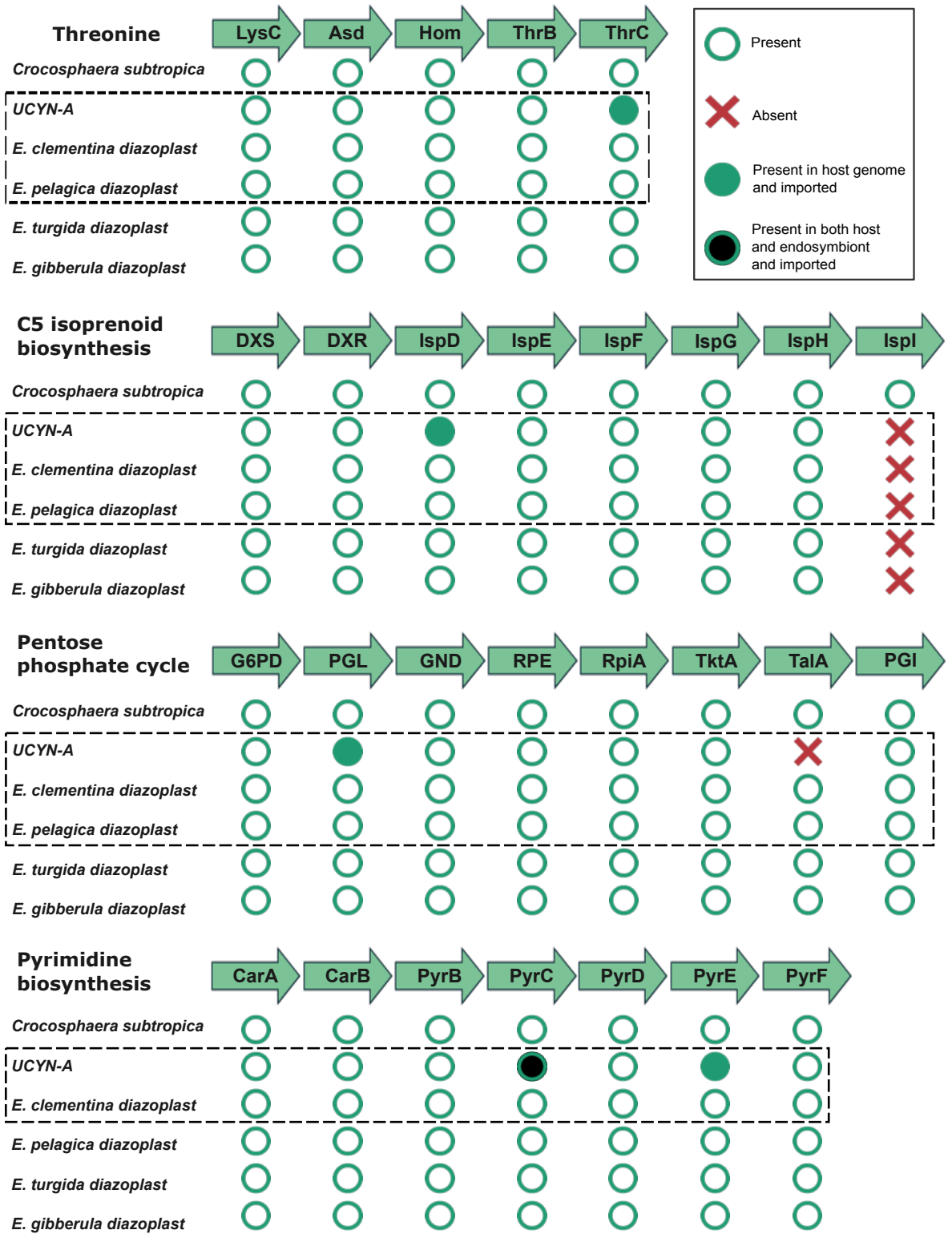
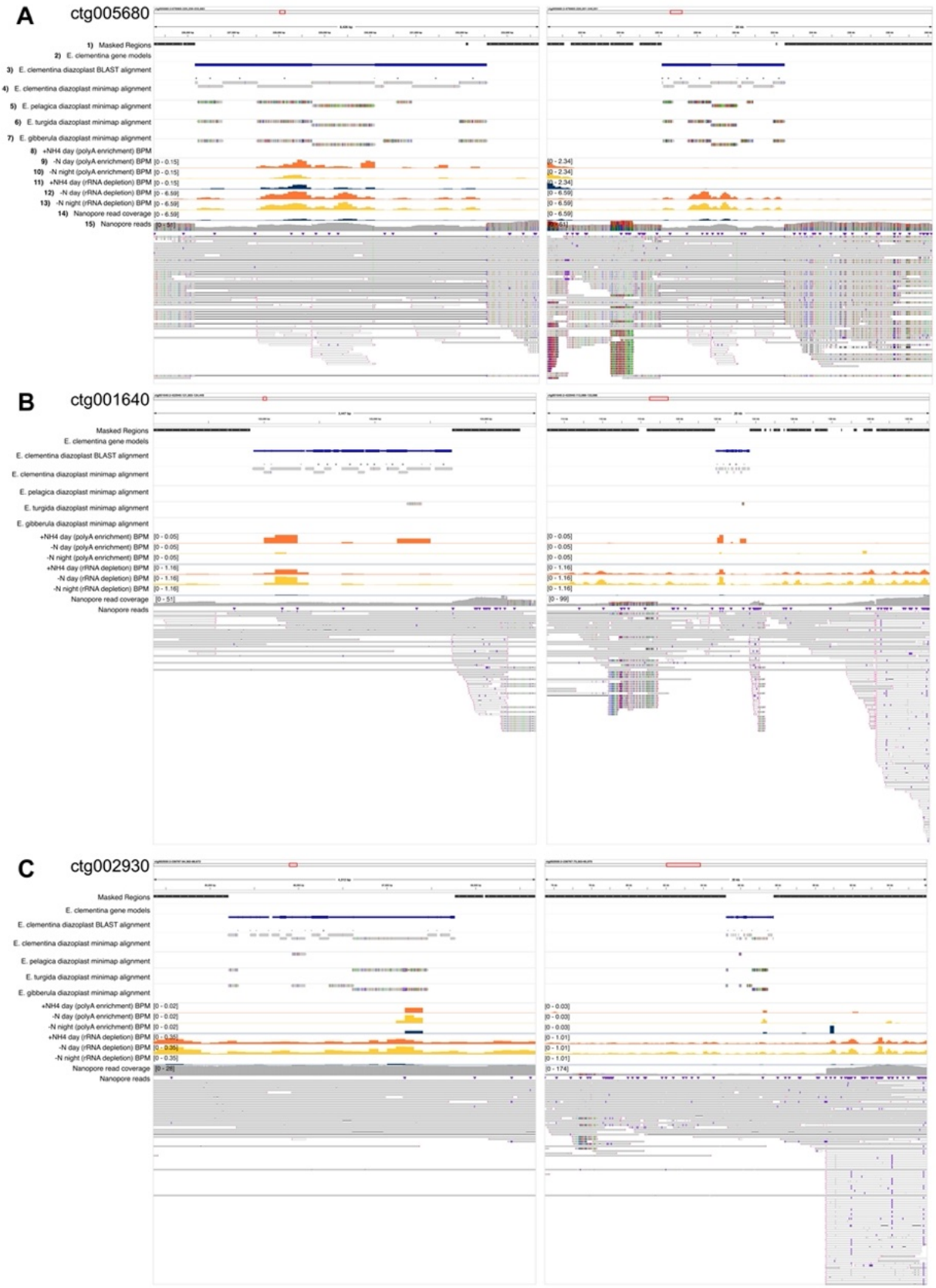
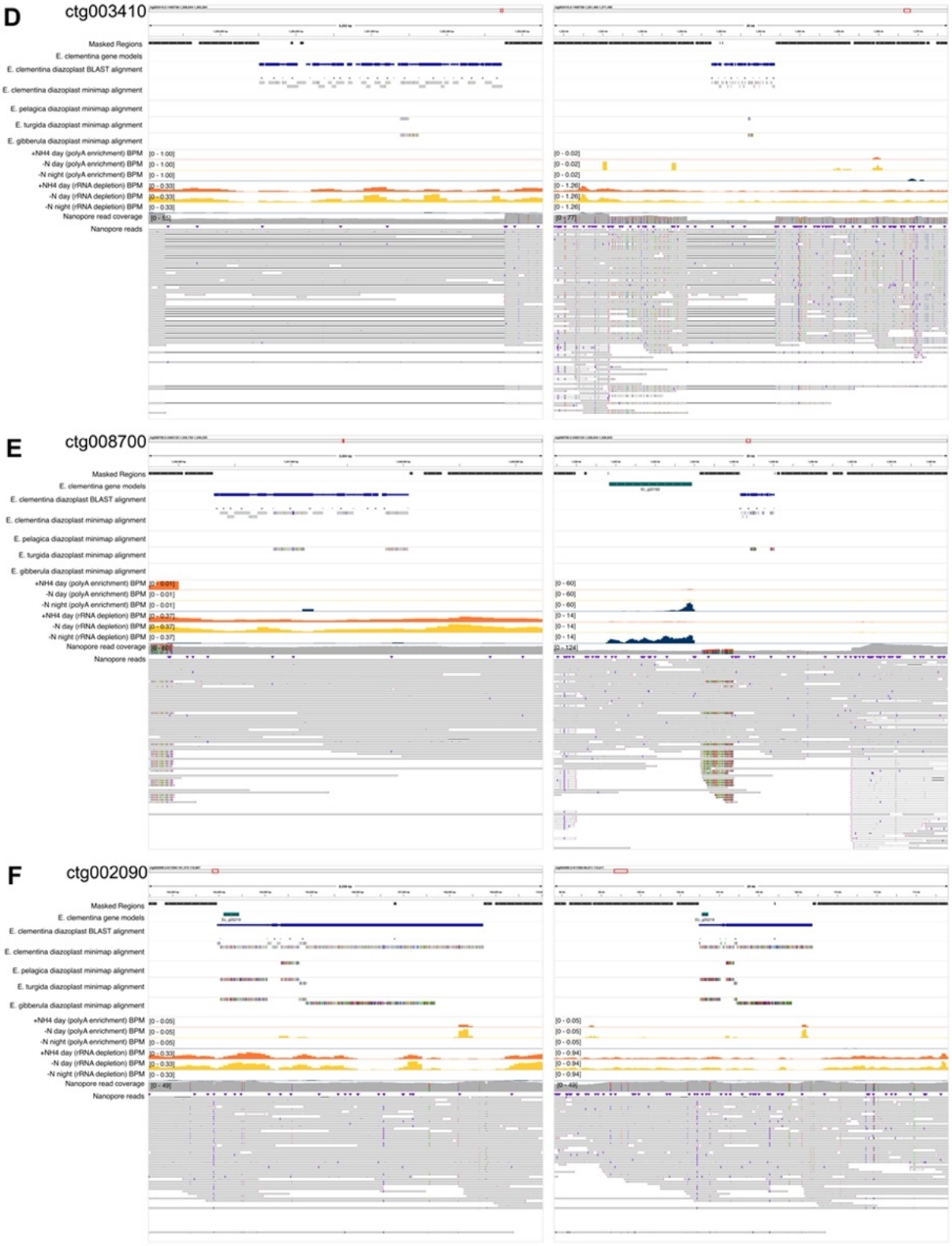
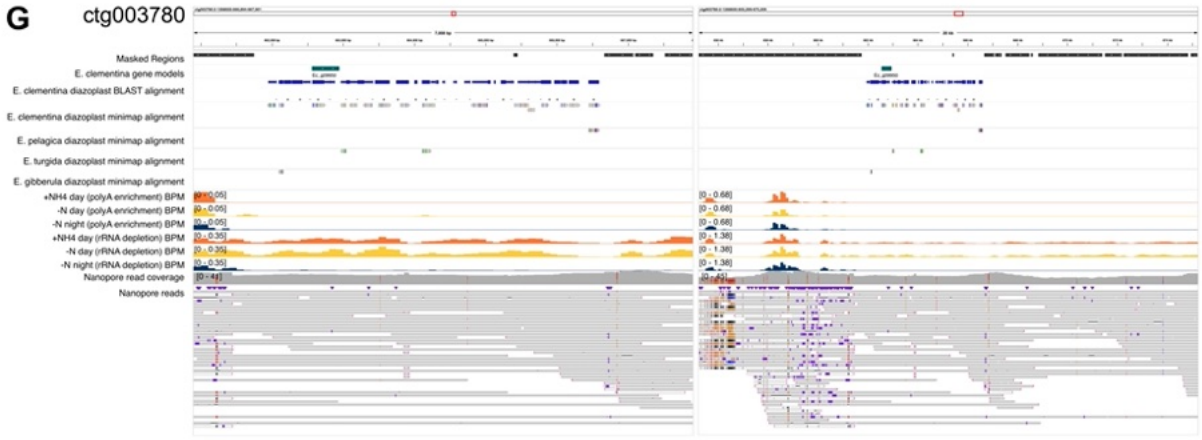


Figure S5. Comparative pathway analysis of diazoplasts and close relatives
KEGG pathway analysis of *E. clementina, E. pelagica E. turgida, E. gibberula* diazoplasts as well as *C. subtropica* and UCYN-A, indicating presence (green circle) or absence (red x) in the genome. Filled green circle indicates evidence for import of a host-encoded protein; filled black circle indicates presence in the endosymbiont genome and evidence for import of a host-encoded protein.







Data S1: Detailed genome tracks across NUDT regions

(A-G) For all NUDTs, full context genome tracks from the Integrated Genomics Viewer zoomed in to the NUDT region (left) or zoomed out to a 20kb surrounding region (right). Tracks from top to bottom are:

1) Region file of masked repeat regions;

2) Feature file of E. clementina gene models;

3) Region file of E. clementina diazoplast homology found by BLAST, demarcates the NUDT;

4-7) Alignment files of homology found by minimap2 when aligning 4) E. clementina diazoplast, 5) E. pelagica diazoplast, 6) E. turgida diazoplast, and 7) E. gibberula diazoplast to the E. clementina nuclear genome;

8-10) Normalized expression data in BPM of RNA seq from combined replicates of poly-adenylated transcript enriched RNA collected from three treatment conditions.

11-13) Normalized expression data in BPM of RNA seq from combined replicates across of ribosomal RNA depleted RNA collected from three treatment conditions.

14) Read pileup of axenic nanopore reads. Colored bars at certain sites indicate proportion of SNVs across the reads deviating from the haplotype reference assembly often resulting from a heterozygous site but sometimes from reads accumulating at undiscernible copies of repeat elements.

15) Alignment file of axenic, nanopore long reads aligned to the reference E. clementina genome. An aligned read identical to the reference sequence is rendered as a single plain grey bar. Colors at sites along the read denote SNVs from the reference assembly. Small indels are denoted by small purple bars. A thin black bar within a read represents a region not present in the read that is present in the haplo-assembly (i.e. larger indels). Very light grey bars are secondary alignments, which accumulate at repeat elements.