You can find four folders and one gzipped file here. These data are the sequence/alignment data analyzed by Sarai et al. (https://www.biorxiv.org/content/10.1101/702274v1), as well as the results (phylogenetic trees) that were omitted from the supplementary information.

- **Folder entitled "Sarai_75protein_analysis".**

This folder contains one sub-folder and an excel file. A sub-folder (single_genes) contains single-gene alignments (75 in total) comprising 75-protein alignment used for a phylogenomic analyses. The alignments were nexus-formatted. The character set called "Gblocks" contains the positions used for phylogenetic analyses. An excel file (Sarai_75protein_alignment_coverage.xlsx) provides the site coverages of individual taxa included in 75-protein alignment.

- **Folder entitled "Sarai_Chla_analyses".**

This folder contains the alignments of 7 enzymes involved in chlorophyll-a biosynthesis (XXX_YYY.phy; XXX and YYY correspond to an enzyme name and amino acid positions included, respectively). All alignment files were trimmed (i.e. ambiguously aligned positions were discarded) and saved as the phylip-format. "XXX_sequencename.tx
t" files (XXX corresponds to an enzyme name) provide the species name from which each sequence in an alignment were originated. The folder also contains a single PDF file of 7 ML phylogenetic trees with ML bootstrap support values.

- **Folder entitled "Sarai_Heme_analyses".**

This folder contains the alignments of 7 enzymes involved in C5 pathway for the heme biosynthesis (XXX_YYY.phy; XXX and YYY correspond to an enzyme name and amino acid positions included, respectively). All alignment files were trimmed (i.e. ambiguously aligned positions were discarded) and saved as the phylip-format. "XXX_sequencename.txt" files (XXX corresponds to an enzyme name) provide the species name from which each sequence in an alignment were originated. The folder also contains a single PDF file of 9 ML phylogenetic trees with ML bootstrap support values.

- **Folder entitled "Sarai_IPP_analyses".**

This folder contains the alignments of 9 enzymes involved in IPP biosynthesis (XXX_YYY.phy; XXX and YYY correspond to an enzyme name and amino acid positions included, respectively). All alignment files were trimmed (i.e. ambiguously aligned positions were discarded) and saved as the phylip-format. "XXX_sequencename.txt" files (XXX corresponds to an enzyme name) provide the species name from which each sequence in an alignment were originated. The folder also contains a single PDF file of 7 ML phylogenetic trees with ML bootstrap support values.

- **Folder entitled "Sarai_PetC_RbcS_PsbO_analyses".**

This folder contains the sets of sequences and datasets used in phylogenetic analyses for RbcS, PsbO and PetC. "XXX_sequences.fasta" files (XXX corresponds to an protein name) provide the sets of sequences used in each analysis in the fasta-format. "XXX_dataset.fasta" files contain the datasets used in the analyses in the fasta-format. All ambiguously aligned positions were removed from the

datasets. The text file “stats.txt” provides the number of sequences and alignment length of each dataset.

- **Nterm_IPP_Heme_Chla_20191217.xlsx**

We summarized the results from the analyses of the N-terminal extensions identified in putative enzymes involved in C5 pathway for the heme biosynthesis, and IPP and Chl-*a* biosynthetic pathways. See comments in the excel file for the details.

- **Assembled_data.tar.gz**

A gzipped file contain the assemble data (fasta-formatted) of MGD and TGD, namely "MRD_trinity.fasta" and "TRD_trinity.fasta."