SI Text

Table of contents

Chromochloris zofingiensis strains and culture conditions

We used the *Chromochloris zofingiensis* strain SAG 211-14 obtained from the Culture Collection of Algae at Goettingen University. The cells were grown at 25°C in liquid cultures shaking at 100–150 rpm in diurnal (16 h light, 8 h dark) conditions with light intensity of 90–100 μmol photons m^{-2} s^{-1} unless stated otherwise. Cells were grown in Proteose Medium (UTEX Culture Collection of Algae) with Chu's micronutrient solution (2 mL/L, UTEX Culture Collection of Algae) unless stated otherwise. Cells were counted with the Multisizer 3 Coulter Counter (Beckman Coulter). Cells were harvested by centrifugation (2,200–4,620 *g* for 5–10 min), discarding the supernatant, resuspending the cells in media and transferring to an eppendorf tube, pelleting by centrifugation (15,000 *g* for 5 min), discarding the supernatant, and freezing the cell pellet in liquid nitrogen unless stated otherwise.

X-ray tomography

Cells were grown until log phase, pelleted by centrifugation (700 *g* for 2 min), and then loaded into custom-made thin-walled glass capillaries (1). Glass capillaries had been previously dipped in a solution of 100 nm gold nanoparticles (EMGC100, BBI International, Cardiff, CF14 5DX, UK), which were subsequently used as fiducial markers for alignment of the X-ray projections. Once loaded into capillaries, cells were cryo-preserved by plunging the tip of the specimen capillary into a ~90 K reservoir of liquid propane at 2 m s^{-1} using a custom-made fast-freezing apparatus.

Soft X-ray tomographic data were acquired using the cryogenic soft X-ray microscope in the National Center for X-ray Tomography (NCXT) at the Advanced Light Source in Berkeley, California. The microscope and image acquisition have been described in detail previously (2, 3). Projection images were collected at 517 eV using a Fresnel zone plate with a resolution of ~50 nm as the objective lens. For each data set, 90 projection images were acquired spanning a range of 180°. During data acquisition, the specimen was kept in a stream of helium gas that had been cooled to liquid nitrogen temperatures to maintain cryo-preservation of the sample. Depending on the thickness of the specimen, exposure times for each projection image varied between 200 and 350 ms. 3-D reconstructions of the X-ray projections were calculated using the software package IMOD after manually tracking fiducial markers on adjacent images for alignment (4). AMIRA (FEI) was used to semi-automatically segment the 3-D volumetric reconstructions into subcellular compartments (lipid droplets, chloroplasts, starch, mitochondria) based on their different gray level ranges. Segmentation of the nucleus was performed manually.

DNA preparation and quality assessment

Genomic DNA was prepared as follows. Total cellular DNA was extracted from cells grown in 1 L cultures to ~5 × 10^6 cells/mL. Harvested cells were resuspended in 300 μL Milli-Q purified water and 500 μL lysis buffer (100 mM Tris-HCl pH 8.0, 40 mM EDTA, 400 mM NaCl, 2% SDS) and incubated for 2 h at 65°C while rotating. 170 μL of 5 M NaCl and 135 μL of 10% w/v CTAB in 700 mM NaCl were added. After incubation for 10 min, the DNA was extracted by adding phenol:chloroform, vigorously shaking, and centrifuging (~15,000 *g* for 5 min) to separate phases. The aqueous phase was removed and placed in a new tube with 5 μL of RNase A, incubated for 20 min at 37°C, and followed by two additional phenol:chloroform extractions and one chloroform extraction. To precipitate the DNA, 0.1× sample volume of 5 M NaCl and 0.7× sample volume of isopropanol were added to the resulting aqueous phase, the sample was mixed, and pelleted by centrifugation (15,000 *g* for 15 min at 4°C). The supernatant was removed and pellet was washed with cold 70% ethanol, centrifuged (15,000 *g* for 5 min at 4°C), and the supernatant removed. The DNA was cleaned with an ethanol precipitation step (100% ethanol, 100 mM sodium acetate pH 5, overnight at 20°C), centrifuged (~15,000 *g* for 5 min at 4°C), and followed by an ethanol wash. The DNA pellet was briefly air-dried and resupsended in Milli-Q purified water. DNA concentration and quality was assessed by optical absorbance on a NanoDrop 2000 spectrophotometer (Thermo Scientific).

To obtain high molecular weight DNA (≈270 Kbp) for optical mapping, 1 L of cells were grown to ~5 × 10^6 cells/ml. The harvested cell pellet was washed twice with cold ethanol and resuspended in buffer (200 mM NaCl, 100 mM EDTA, 10 mM Tris, pH 7.2). An equal volume of 1% agarose was gently mixed with the cells and the cell-agarose suspension was aliquoted into plug molds and cooled (4°C for ~60 min). The cell wall was digested by incubating the cell plugs in protoplasting solution (4% w/v hemicellulose, 2% w/v driselase, 1 M sorbitol, 5 mM sodium citrate, 240 mM EDTA pH 8.0, 10 mM 2-mercaptoethanol) overnight at 37°C while shaking. To lyse the cells, the protoplasting solution was removed and the cell plugs were incubated in lysis solution (0.5 M EDTA pH 9.5, 1% w/v *N*-lauroylsarcosine, 5 mg/ml proteinase K) overnight at 50°C. The lysis solution was removed and the cell plugs placed in 0.5 M EDTA pH 9.5 and shipped to OpGen, Inc. for optical mapping using BamHI enzyme.

RNA preparation and quality assessment

RNA was prepared as follows. Cells were washed with cold ethanol on dry ice and ethanol was removed by centrifugation (2,200 *g* for 3 min at 4°C). To break cells open, cells were homogenized with lysing matrix D on dry ice for 2× 60 s with the FastPrep-24 (6.0 m s^{-1}, MP Biomedicals). Buffer (50 mM Tris-HCl pH 8.0, 200 mM NaCl, 20 mM EDTA, 2% SDS, 1 mg/mL proteinase K) was added, samples were vortexed and incubated for 3 min at room temperature, and cell debris was pelleted by centrifugation (20,000 *g* for 3 min). 1 mL of sample was added to 10 mL of TRIzol in MaXtract HD tube and incubated for 3 min at room temperature. To extract RNA, 1/5 volume chloroform was added, samples were vigorously shaken, incubated for 5 min at room temperature, and phases were separated by centrifugation (800 *g* for 5 min at 22°C) and decanting. Total RNA was precipitated by adding cold ethanol on the aqueous phase and purified using the miRNeasy mini kit (Qiagen). RNA was eluted with DEPC-treated water and cleaned with an ethanol precipitation step (100% ethanol, 85 mM sodium acetate pH 8.0), centrifugation (~15,000 *g* for 5 min at 4°C), and ethanol washing. The pellet was briefly air-dried and resuspended in DEPC-treated water. RNA concentration and integrity was assessed by NanoDrop 2000 spectrophotometer (Thermo Scientific) and Agilent 2100 Bioanalyzer.

RNA-Seq

Total RNA was purified from each culture as described above. The rRNA was selectively depleted with the Ribo-Zero rRNA Removal Kit (Plant Leaf) according to the manufacturer's instructions (Illumina). The remaining RNA was converted into cDNA and made into sequence-ready libraries with the KAPA Stranded RNA-Seq Kit (Kapa Biosystems). The 14 *de novo* transcriptome RNA-Seq libraries were pooled and sequenced with 150+150 bp paired-end reads on two lanes of a HiSeq 2500 high-throughput sequencer according to manufacturer's instructions (Illumina). The 44 high light RNA-Seq libraries were combined into three pools and sequenced with 50 bp single-end reads on three lanes of a HiSeq 2500.

The resulting data was demultiplexed with in-house scripts. Adapter sequences were trimmed with Scythe (5) and aligned to the ChrZofV5 release of the *C. zofingiensis* genome with RNA STAR (6). Determination of counts per gene and transcript abundance in terms of fragments per Kbp of gene per million mapped fragments (FPKMs, Datasets S20–S21) were made with Cuffdiff (7). Further analyses and figures were generated with cummeRbund package in the R statistical computing environment (8). PCA was performed with `plotPCA()` from the R affy package (9). Two-fold differentially-expressed genes and regularized $\log_2$-transformation were performed with the R DESeq2 package (10).

De novo transcriptome conditions

Transcriptome material was derived from 100 mL cultures of cells (~4–9 × 10^6 cells/mL) from 14 different conditions: high light (400 µmol photons m^{-2} s^{-1}), medium light (100 µmol photons m^{-2} s^{-1}), low light (10 µmol photons m^{-2} s^{-1}), glucose (20 mM), 48 h darkness, 4 h anaerobic, 4 h dark and anaerobic, 1 h without sulfur (Bristol's Medium without $MgSO_4$, UTEX Culture Collection of Algae), 1 h without nitrogen (Bristol's Medium without $NaNO_3$), 1 h without phosphorus (Bristol's Medium without K_2HPO_4, KH_2PO_4), 1 h without iron (Bristol's Medium), low oxidative stress (5 µM rose bengal, 0.5 h dark followed by 1 h 100 µmol photons m^{-2} s^{-1}), high oxidative stress (5 µM rose bengal, 0.5 h dark followed by h 100 µmol photons m^{-2} s^{-1}), and hydrogen peroxide oxidative stress (1 mM H_2O_2). Cells were collected by centrifugation (2,200 *g* for 5 min at 4°C), the supernatant was discarded and the cell pellet was frozen in liquid nitrogen.

Changes in gene expression during shift to high light

The gene expression light intensity experiment from medium light (100 µmol photons m^{-2} s^{-1}) to high light (400 µmol photons m^{-2} s^{-1}) was conducted as follows. 1 L cell cultures were grown to log phase (~3.0 × 10^6 cells/mL) under medium light (100 µmol photons m^{-2} s^{-1}). Cultures were mixed and divided into 75 mL cultures in sterile 250 mL beakers. After acclimating overnight, the light treatment cultures were moved from 100 µmol photons m^{-2} s^{-1} to 400 µmol photons m^{-2} s^{-1}, while control cultures were maintained under 100 µmol photons m^{-2} s^{-1}. Replicates ($N = 4$) were collected at 0, 0.5, 1, 3, 6, and 12 h, harvested by centrifugation (200 *g* for 5 min at 4°C), and frozen in liquid nitrogen. RNA was extracted, processed, and analyzed as described above.

Assembly overview

Next-generation sequencing and associated software has made draft assemblies via short-read whole genome shotgun sequencing easy and relatively automatic. For eukaryotic organisms, these drafts are typically highly fragmentary by traditional standards of model organisms, with fragments often of size spanning only one to a few genes at a time. For *Chromochloris*, we aimed for a chromosome-level assembly comparable to model organisms, and initial drafts purely via automated short-read methods were only of "gene-space" quality and did not meet the goal. Hence, additional data — a global optical restriction fragment map from OpGen, Inc. and long reads via Pacific Biosystems ("PacBio") — were collected. No software was found able to automatically incorporate this additional data well enough to meet the assembly goal; hence, extensive manual integration effort was expended to meet the goal starting from automated assemblies as a base. As the methods used are uncommon, they are described in detail below.

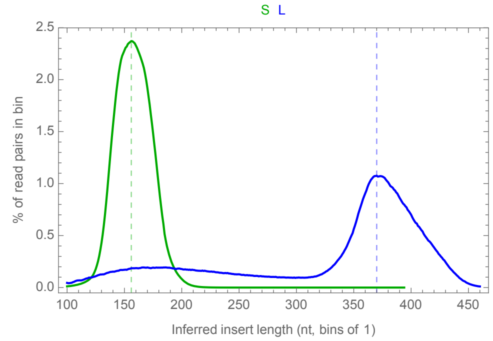
Genomic and RNA-Seq sequences

Two Illumina paired-end libraries — "S" with shorter and "L" with longer inserts — were prepared as described earlier for genomic (combined nuclear, chloroplast, and mitochondrion) sequencing, including Illumina inline controls and a small amount of Illumina PhiX. Each library was run as an entire single lane of a HiSeq 2000 V3 flowcell at the UCLA BSCRC Sequencing Core to obtain ~104M ("S") and ~66M ("L") paired end 100+100 nt reads with ~96% of pairs passing RTA PF=1 (PF=0 pairs were discarded). (Pacific Biosciences genomic reads are discussed later.) Fourteen Illumina TruSeq paired-end RNA-Seq sub-libraries were prepared as described earlier. A single equi-molar pool was run on both lanes of a HiSeq 2500 V1 rapid flowcell at the UCLA BSCRC Sequencing Core to obtain ~476M 151+151 nt read pairs with 7 nt TruSeq index reads with ~86% of pairs passing RTA PF=1 (PF=0 pairs were discarded). Demultiplexing for assembly by perfect match to expected 7-mers gave ~23M to ~34M read pairs per sub-library and ~397M (~97% of PF=1) read pairs total.

Analyses of reads, a multitude of *in silico*-targeted subsets of reads, and various fractions of reads (e.g., heads or tails of first or second ends) were made over many iterations, starting with exploratory preliminary analyses under minimal assumptions and proceeding toward final analyses as conclusions and partial results accumulated. Tools used included assemblers Ray (11), ABySS (12), and ALLPATHS-LG (13, 14); aligners Bowtie (15), Bowtie2 (16), HISAT/HISAT2 (17), BLAST (18), BLAST+ (19), LAST (20), LASTZ (21), BLAT (22), OpGen, Inc.'s MapSolver, BLASR (23), and Parasail (24); error correctors / double-sequenced end overlappers / adapter trimmers Proovread (25), SeqPrep (26), and Cutadapt (27); sequence analyzers Jellyfish (28), MUMmer (29), TRF (30), IRF (31), RepeatMasker (32) with Repbase Update (33), and RepeatModeler (34); gene callers AUGUSTUS (35) and tRNAscan-SE (36); visualization/analysis tools Savant (37), IGB (38), IGV (39), Biomatters Limited's Geneious, and Circos (40); GUI automaton Keyboard Maestro of Stairways Software Pty Ltd.; standard UNIX text-processing tools as well as bioinformatic utilities such as SAMtools (41), DEXTRACTOR (42), HTSeq (43), and EMBOSS (44); databases of biological knowledge such as NCBI (45), Pfam (46), and Rfam (47); as well as custom one-off programs and scripts written in languages such as C++, Perl, Wolfram's Mathematica, and MathWorks's MATLAB. Some computations were carried out on the UCLA Hoffman2 computing cluster.

Insert lengths and read preparation/composition

From preliminary and later assemblies, mode insert lengths exclusive of adapters were ≈156 nt for "S" and ≈370 nt for "L", with "S" fairly Gaussian with standard deviation ≈16 nt, but "L" bimodal with approximately one third in a wide mode at ≈200 nt and two thirds in a non-Gaussian narrower mode at ≈370 nt skewed longer.



Inserts below 100 nt read into adapters (empirically verified to be as expected: for "S", first end `A` + TruSeq #6 + dark/poly-`A`, second end reverse complement of TruSeq universal adapter + dark/poly-`A`; for "L", same except with TruSeq #12). Little of the "S" and "L" distributions is so short, and only ~150K (< ~0.2%) of "S" and ~271K (< ~0.5%) of "L" pairs contain ≥ 1 16-mer of consensus adapter ignoring dark/poly-`A` tails. Preliminary analyses often did not try to identify and remove adapters, while later analyses generally had them stripped via SeqPrep or Cutadapt.

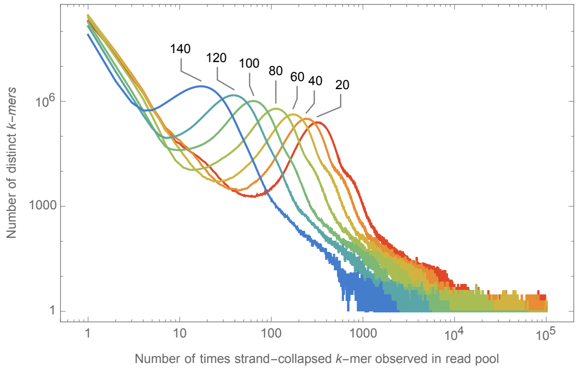
Inserts below 200 nt have overlapping ends: almost all of the "S" distribution is as such, and a fraction of the shorter "L" mode is as well. Early analyses identified overlapped ends (merging double-sequencing to form consensus virtual single end reads) via unique overlaps of ≥ 16 nt with ≤ 3 mismatches; ~88% of "S" pairs and ~7% of "L" pairs were merged. The resulting pool of reads used for initial assemblies was then ~89M and ~5M virtual single end reads of total sizes ~14 Gnt and ~0.7 Gnt, and ~12M and ~59M read pairs of total sizes ~2 Gnt and ~12 Gnt, for a grand total of ~28.6 Gnt. Later analyses used SeqPrep for overlap detection and merging.

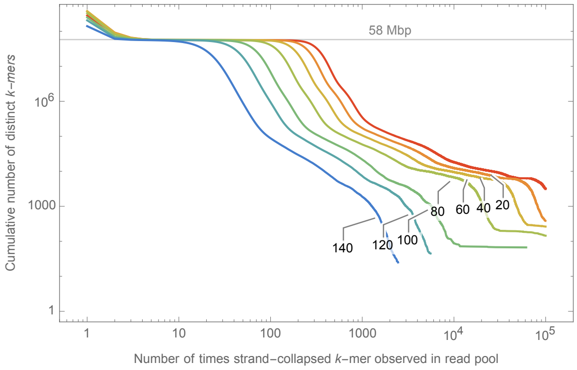
Rough composition is ≈1.3%/1.0% of "S"/"L" pairs as Illumina inline process controls (with ~95%/97% of pairs with ≥ 20% of 16-mers hitting a known control having ≥ 80% of 16-mers being hits) and ≈1.8%/2.9% as PhiX (with ~94%98% of pairs with ≥ 5% of 16-mers hitting *de novo* circular PhiX having ≥ 2/3 of 16-mers being hits), leaving ≈97%/96% for nuclear genome + chloroplast + mitochondrion. Once organelle genomes became available,, ≈0.5%/0.7% and ≈0.2%/0.2% was estimated as chloroplast and mitochondrion, respectively.

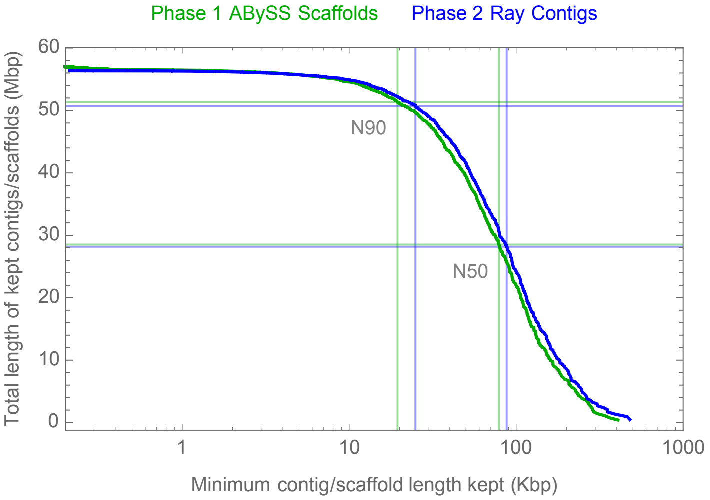
With (1) coverage plentiful relative to the ≈58 Mbp assembly size estimate (see next section), (2) ≈70%/80% of "S"/"L" PhiX read pairs manifestly error-free, and (3) several dozen not unlikely corruption possibilities of comparable probability existing for a typical read (e.g., although PhiX errors concentrated as expected at the tails of reads, error position probability was substantial across more than 20 nt), it was decided to not generally perform spectral-based read "error correction" procedures on the Illumina reads. (However, as discussed later, correction of the Pacific Biosciences reads was critical for their use in refining the assembly.

Nuclear assembly Phases 1 and 2: Automated base assemblies

Histograms of the number of times distinct strand-collapsed (e.g., Jellyfish "canonical") *k*-mers appeared in the prepared Illumina read pool for various *k* suggested that potential diploidy was not a great concern, multi-copy repeats (although surely present) did not constitute an excessive fraction of the genome, and there were no large contaminants (e.g., bacterial genomes), suggestions later supported by data such as the ~58 Mbp optical size estimate and BLAST comparisons of final genome products against the universe of NCBI sequences. Plateaus visible in the cumulative plot provided one of the filterable signals by which the assembly of non-chromosomal sequences (discussed later) began.





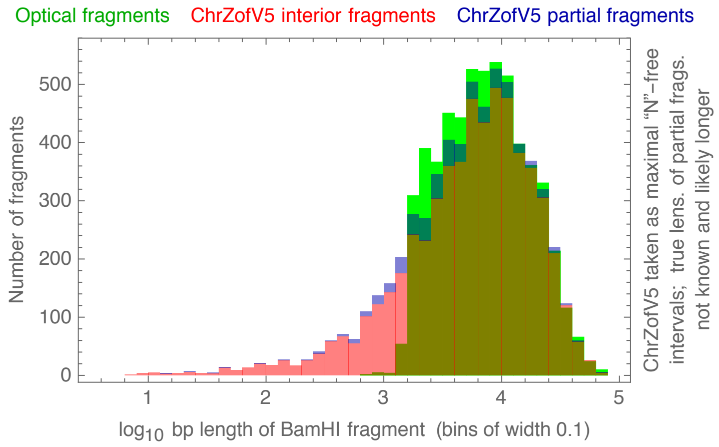


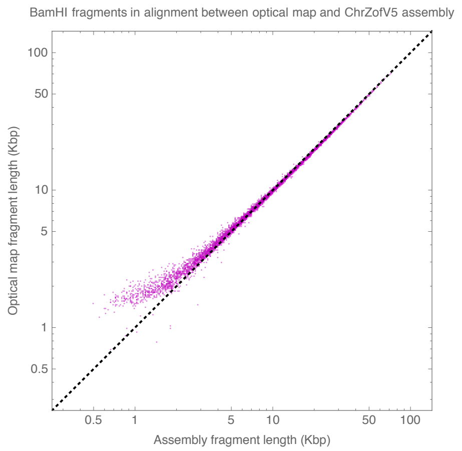
The main automated draft assembly used in the first years of the project ("Phase 1") was an ABySS *k*=95 "gene-space" one on the prepared Illumina reads consisting of 3,513 scaffolds with longest ~407 Kbp, N50 ~79 Kbp, N90 ~19 Kbp, L50=217, and L90=754. This assembly guided further decisions (e.g., use of optical mapping and with BamHI) and was the point at which downstream analyses such as gene prediction began. Because Phase 1 contigs were slightly shorter than needed for high-likelihood automatic optical map placement, in Phase 2 additional assemblers were tried in an effort to find a slightly better automated base; a Ray *k*=51 "gene-space" one on the prepared Illumina reads consisting of 1,335 contigs with longest ~479 Kbp, N50 ~88 Kbp, N90 ~25 Kbp, L50=193, and L90=652 was chosen.

Nuclear assembly Phase 3: optical map and chromosome-level scaffolding, joining, filling

OpGen, Inc. was contracted to construct an "optical map" of *Chromochloris* by imaging immobilized complete restriction digests of linearly-combed large molecular weight pieces ("hunks") of genomic DNA we provided. Based on Phase 1, they chose BamHI (`G|GATCC`) as digest enzyme due to the range of predicted fragment lengths being mostly accessible to their technology. They ran 12 high-density MapCards to obtain approximate fragment length fingerprints for ~318K hunks, which they assembled into 19 maptig chromosomes of total size ~58 Mbp whose constitutents are not `A`/`C`/`G`/`T` nucleotide calls, but approximate fragment lengths under complete BamHI digestion (Datasets S6–S7). (As they omit all small maptigs, chloroplast and mitochondrion do not appear.) The final nuclear genome nucleotide sequences described in this work — the "ChrZofV5" ver. 5 assembly (Datasets S1–S3) that Phase 4 (described later) ends with — adopt chromosome numbering and '+' strand decisions from this optical assembly.

During optical assembly, hunk fingerprints are piled up in multiple alignments with typical coverage of several dozens; chromosome ends are manifest as consensus locations beyond which hunk fingerprints do not extend (up to uncertainty in optically-estimated fragment lengths). OpGen observed both ends of all chromosomes except the right end of chromosome 5, the tail of which assembled into an approximate optical inverted repeat that, as discussed further later, likely is just the beginning of a much longer true sequence inverted repeat. (Due to this, the optical length of chromosome 5 is likely underestimated by ~0.56 Mbp and chromosome numbering does not reflect true size largest to smallest.)





OpGen's MapSolver software visualizes the optical map and aligns sequence contigs/scaffolds to it. (A degree of mismatch is allowed due to optical length uncertainty, the tendency of small fragments to be lost optically, and the possibility of small basecall sequence errors creating or deleting cutsites.) Experience suggests a contig/scaffold needs ≥ 5 interior fragments of non-small length (≥ ≈2 Kbp) for MapSolver to have a reasonable probability of placing it. This translates into a wide variety of contig/scaffold lengths due to *Chromochloris* BamHI fragment size variation, and Phase 1 scaffolds were often near this threshold with only ~12% covering ~37.6 Mbp being automatically placeable, even allowing non-unique placements and multiple coverage; the slightly longer contigs of Phase 2 improved to ~29% covering ~38.2 Mbp. However, by Phase 4's end with extensive hand work, ~93% of the optical map was uniquely covered (see main text Fig. 2) with just a single sequence scaffold per optical chromosome.

Most automated assemblers have as a design goal to be conservative, in that they would prefer to give a more fragmented result (which could be pasted together in an unknown way to get "truth") rather than one with mis-assemblies (in which some contigs/scaffolds would need to be taken apart before pasting could arrive at truth). Consistent with this, only a handful of Phase 1/2 scaffolds/contigs were found to be mis-assemblies via alignment to the optical map, increasing confidence that base sequence at finer resolution than the optical map was generally correct.

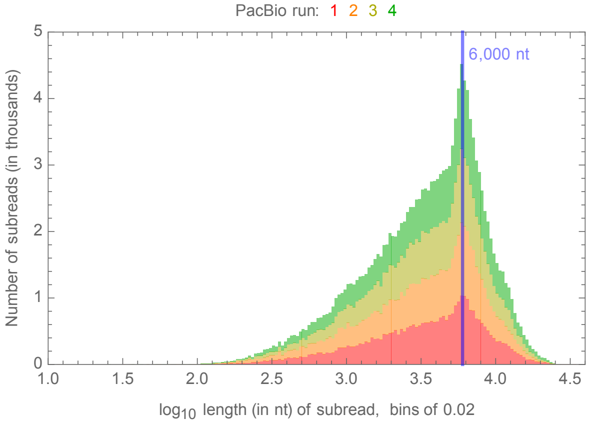
The per-chromosome single sequence scaffolds were formed from iterative rounds of optical placements of smaller subsequences (longer and longer as hand work proceeded), with the optical map providing global, externally-validated subsequence ordering and strand orientations and enabling approximate but accurately-sized `N`-filled gaps among subsequences and chromosome edges. Placements that resulted in overlapping or touching subsequences up to optical length uncertainty were, e.g., inspected at the sequence level for nucleotide overlap agreement of shorter lengths than automatic assemblers might otherwise require; reads and read pairs (including Pacific Biosystems long reads once Phase 4 began) touching and spanning gaps were isolated; and ambiguous placements could sometimes be resolved in favor of those not covering already well-covered parts of the optical map.

The optical map was intensely useful: as hand work proceeded, speculative contig joins and extensions that would have been dangerous — likely forming mis-assemblies if too many were relied upon, especially in succession — became reliable, as once further BamHI sites were reached, independent verification by the optical map was attained and possibilities were eliminated. Similar to the physical maps used in model organism projects, the optical map provided a global ground truth and acted as a ratchet for making positive progress that kept hand work from compounding mistakes. The flavor became much like a jigsaw puzzle, with each additional placement generally making other placements easier, as one could focus on gaps and not only were gaps getting fewer and smaller, but the pile of contigs, scaffolds, and reads to fill them with was also shrinking. Sources for speculation included: alignments of contigs and scaffolds to themselves; re-alignments of reads and read pairs to contigs and scaffolds; and, most importantly, the Pacific Biosciences long reads of Phase 4. Consensuses of supporting evidence spanning gaps was used to fill gaps; in some cases, these gap fills are of low quality (e.g., naked single-read PacBio sequence) but it was felt that — as long as the evidence supporting the join was substantial — it was better to provide some representative sequence and close gaps rather than fret for first public release over every basepair being absolutely certain.

Nuclear assembly Phase 4: Pacific Biosciences reads, contig joining, and gap filling

Numerous barriers in the draft assemblies evidently arose from the use of only short, paired end Illumina reads. Many difficulties were near repetitive sequence, either: (1) non-short segments of moderate/high entropy DNA that occur multiple times in the genome; or (2) low entropy DNA (e.g., microsatellites), these being trouble because of either (a) ambiguous continuations due to only having short reads, or (b) coverage collapse of multiple orders of magnitude (even so far as to completely deplete our nearly half-a-thousand-fold average coverage). Difficulties of type (2b) were common at points of very high G+C content as short as a dozen or two basepairs (as has been the authors' experience on other projects with the Illumina platform), and there appear to be many such loci in this genome (ChrZofV5 has 191 clusters of G+C runs of length ≥ 24 nt).

To overcome some of these obstacles and better scaffold subsequences, four 75 fps 3 h PacBio RS-II/Springfield 1.1 runs of genomic DNA with BluePippin selection were performed at the DNA Sequencing & Genotyping Center of the Delaware Biotechnology Institute to obtain long reads, but of relatively low quality. Each SMRT cell contained 163,482 ZMWs ("wells"). Of wells with ≥ 1 insert called by the PacBio basecaller, ~94% had only a single insert; hence, only a single longest interval per well was retained from the intersection of the "insert" and "HQ" regions, and no circular consensuses were made. The result was 149,364 "subreads" (from ~30% of wells) of 12 nt to ~34 Knt (median ~3.1 Knt) of total length ~692 Mbp.



As usual, per-base error rates were estimated by the PacBio basecaller as very high compared to Illumina reads and as mostly indels rather than substitutions. No subread basecalls were given combined substitute/insert/delete/merge Phred qualities ≥ 15 (~3% chance of error or better), the mode was Phred 13 (~5% error), ~25% had Phred quality 0 to 7 (~20% to ~100% error), and the average chance of error per base was ~18%. Hence, pre-correction alignments of subreads to assemblies used PacBio-aware BLASR. While each default 12-mer seed only has ≈9% chance of being uncorrupted, queries ≥ ≈220 nt long have estimated chance ≥ ≈99% of ≥ 1 uncorrupted seed. Most BLASR parameters were left at defaults, but filtering was lowered to impose no minimum read/subread length and no percent identity requirement, and best alignments per query was raised to 250 internal and 100 emit as (1) the shortest target contigs were ≈200 bp and longest PacBio reads ≈50x longer; and (2) alignments descending into false positives were desired so that their statistics could be inferred from their great numbers.

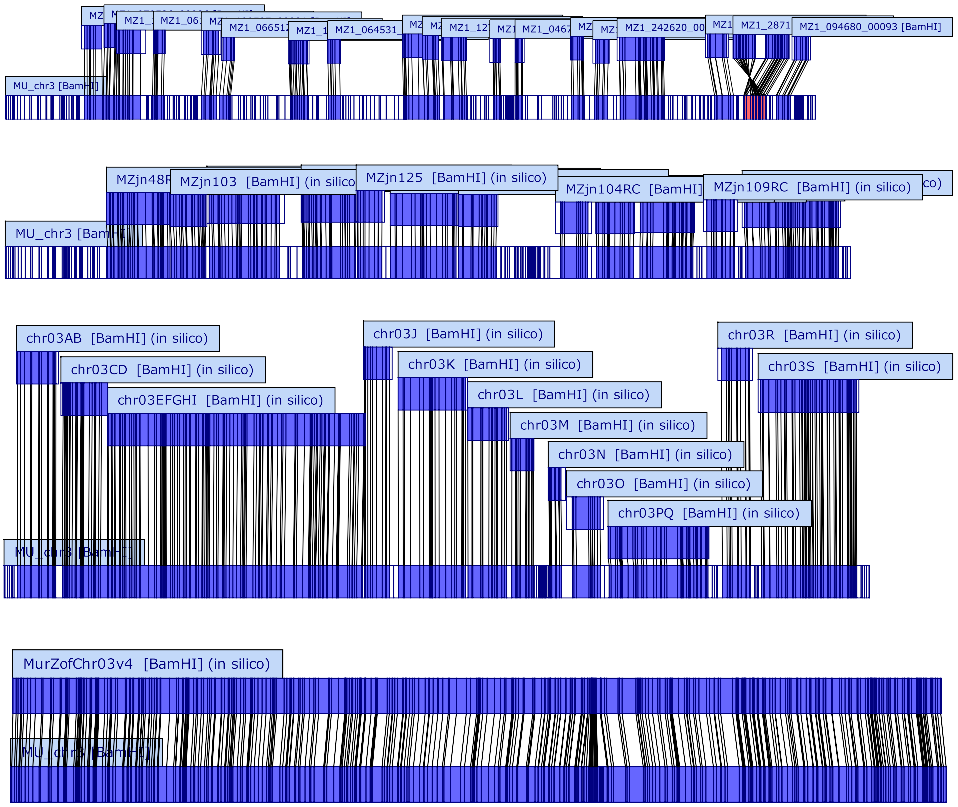
From histograms of query and target alignment spans, a threshold of ≥ 140 nt was chosen for both spans to separate most true hits from very short false positives as well as short sequences repetitively occuring in *Chromochloris*, resulting in ≈0.5M alignments. (Repetitive sequences of longer lengths remained; pre-alignment masking by tandem repeat finder TRF was sometimes used to help.) In subreads with ≥ 1 alignment, ~88% / ~7% / ~5% of PacBio bases on average participated in exactly one / zero / multiple alignments. For draft contigs participating in ≥ 1 alignment, mode coverage by PacBio bases was typically 8, with ≈0.1% / ≈0.6% / ≈97% of bases uncovered / covered exactly once / covered 2 to 23 times. PacBio reads did not show nearly as much coverage variation across sequence the Illumina platform found difficult (e.g., at runs of G+C's). The top alignment by BLASR score per subread was enriched for near-full length alignment span on the query. Once organelle genomes became available (discussed later), estimates put ≈0.1% / ≈0.2% of aligning subreads as mitochondrion / chloroplast.

PacBio subread alignments were repeatedly used to help make assembly subsequence joins and to fill gaps as mentioned in Phase 3. A typical pass began by extrapolation of the unaligned ends of each aligned subread by the average compression/expansion ratio from indels in the aligned portions. Extrapolations might ("overhang") or might not extend beyond a subsequence's boundary, but overhangs of ≥ 1 Knt were not uncommon and, similar to earlier filtering, those of ≥ 140 nt were deemed "interesting". Based on histograms of distance of alignment starts and stops to subsequence edges for subreads with interesting overhangs, it was decided to consider alignment starts and stops within 9 bp of a subsequence's edge as having reached the edge. A subread alignment with an interesting overhang to a subsequence was considered "linkable" if it reached the same end of the subsequence (both ends for those with interesting overhangs on both ends). Subreads with a single linkable alignment on each end and to different subsequences on each end were declared "linking"; each of these suggests a merging of two subsequences with a particular relative distance and orientation with explicit sequence to fill any gap. Suggested merges from linking subreads were collected into a directed graph (that was typically enriched for linear paths) and evidence weighed at nodes with multiple incident arcs to determine if one arc had much more support (e.g., 6-fold more) than others, in which case only the dominant arc was retained and otherwise all arcs removed. The resulting directed graph of linear chains provides a round of up to a few hundred tentative assembly subsequence joins and gap fills to participate in the hand work process discussed in Phase 3. It was always satisfying to merge two or several subsequences into a subsequence large enough that optical placement became probable, and then finding the new subsequence had a unique optical placement that perfectly filled a hole in existing placements.

Using the pool of prepared but unassembled Illumina short reads as reference, the Proovread error corrector was also run on PacBio reads from three of the SMRT cells to obtain 83,069 polished trimmed reads of total length ~292 Mnt whose lengths were primarily between 500 nt and ~21 Knt (median ~3.0 Knt), with almost all bases explicit `A`/`C`/`G`/`T`s (rare isolated `N`s) and almost all per-base Phred-scale quality scores ≥ 19 (≈1 in 79 chance of error or better). These were very useful, as they enabled use of non-PacBio-aware tools (BLAST, …) to query and manipulate the long read dataset, and were used both in ways similar to the uncorrected reads (e.g., in procedures like the previous paragraph) as well as more targeted questions that arose once two subsequences were placed near each other on the optical map. (During operations such as subsequence joining, the larger pile of untrimmed corrected reads also produced by Proovread was queried as well; in certain cases, this was the only way to make progress and gap fill exposes naked single-read PacBio sequence.)

Periodically, and one last time at the end of Phase 4, prepared Illumina reads not aligning to the working assembly were re-*de novo* assembled to maintain an accurate pool of unplaced contigs/scaffolds. Only those of length ≥ 1 Kbp with less than one third of their 31-mers already represented were retained for the final `chrUn#####` unplaced contigs/scaffolds in the ChrZofV5 assembly release. To simplify naming, a few had a small number of `N`s suffixed to make all their lengths unique.

As an example of the progress made during hand work, the top of the next page shows snapshots of chromosome 3 optical placements at four intermediate stages, from near the beginning of Phase 3 to near the conclusion of Phase 4.


MZ1_066512
MZ1_064531
MZ1_242620_00
MZ1_2871
MZ1_094680_00093 [BamHI]
MU_chr3 [BamHI]
MZjn48R
MZjn103 [BamHI] (in silico)
MZjn125 [BamHI] (in silico)
MZjn104RC [BamHI]
MZjn109RC [BamHI] (in silico)
MU_chr3 [BamHI]
chr03AB [BamHI] (in silico)
chr03CD [BamHI] (in silico)
chr03EFGHI [BamHI] (in silico)
chr03J [BamHI] (in silico)
chr03K [BamHI] (in silico)
chr03L [BamHI] (in silico)
chr03M [BamHI] (in silico)
chr03N [BamHI] (in silico)
chr03O [BamHI] (in silico)
chr03PQ [BamHI] (in silico)
chr03R [BamHI] (in silico)
chr03S [BamHI] (in silico)
MU_chr3 [BamHI]
MurZofChr03v4 [BamHI] (in silico)
MU_chr3 [BamHI]

Overall structure of the nuclear genome

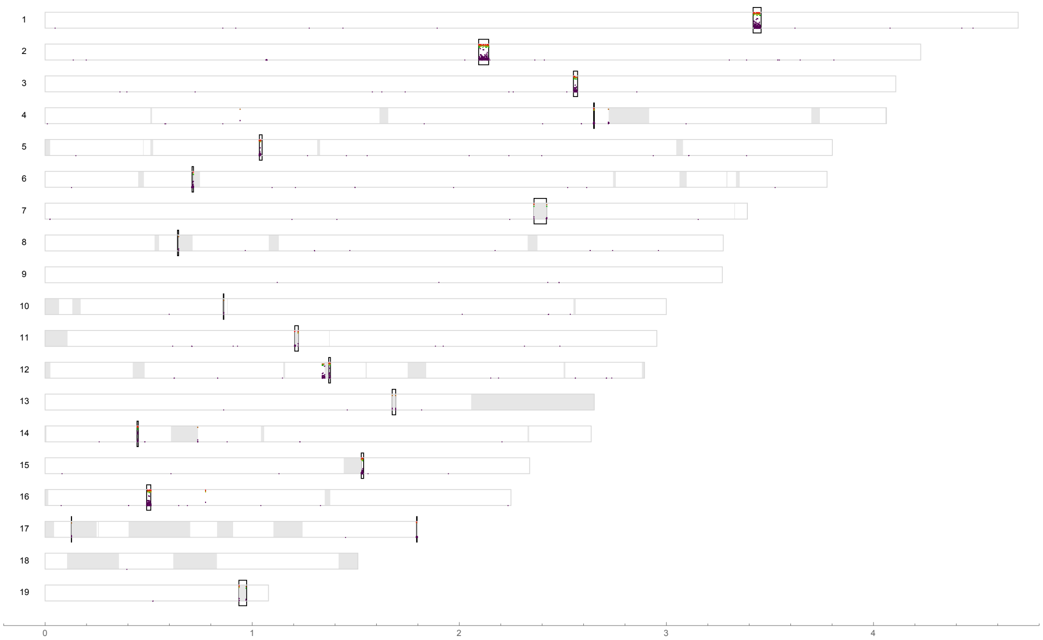
Telomeres. As Phases 3 and 4 progressed and chromosome-level contigs/scaffolds approached optical ends of a chromosome, junctions with telomere repeats became apparent, and efforts were made (returning to Illumina and PacBio reads as necessary) to extend all sequences near such junctions at least partially beyond the junctions. As evident from chromosomes 1–4, 6–9, 13, 15, and 18–19 of the final ChrZofV5 assembly, the canonical *Chromochloris* telomeric repeat is apparently $(\texttt{CCCTAAA})_n$ at 5′-ends of chromosome strands, and from chromosomes 1–3, 6, 8–11, 14–17, and 19 is $(\texttt{TTTAGGG})_n$, the reverse complement, at 3′-ends. From examination of edges of assembly sequences from the algal genomes of Table S1, *Coccomyxa* and *Chlorella* and possibly *Monoraphidium* are the same as *Chromochloris*, although *Chlamydomonas* appears to use $(\texttt{CCCTAAAA})_n$ and $(\texttt{TTTTAGGG})_n$. In *Chromochloris*, commonly observed non-canonical units are $(\texttt{CCTAAAA})_n$ and $(\texttt{CCCTGAA})_n$ near 5′-ends, and $(\texttt{TTTTAGG})_n$ and $(\texttt{TTCAGGG})_n$ near 3′-ends.

A prepared pool of Illumina reads was aligned with Bowtie2 in single end mode keeping top hit only to the ChrZofV5 assembly with PhiX; parameters were end-to-end "`--sensitive`" defaults, which allow short indels and up to ~10% mismatches. Total pool nucleotides aligning to nuclear components was ~26.8 Gnt, and the total size of pool members with ≥ 2 adjacent copies (not necessarily the same) of TAAACCC, TAAAACC, or TGAACCC or ≥ 2 adjacent copies (not necessarily the same) of AGGGTTT, AGGTTTT, or AGGGTTC was ~62 Mnt. As the nuclear genome is ≈57 Mbp, this suggests *Chromochloris* telomeres total ≈133 Kbp (≈3.5 Kbp/end).

The beginning (relative to nominal '+' strands) of chromosomes 5, 10–12, 14, and 16–17 and the end of chromosomes 4–5, 7, 12–13, and 18 were not reached in ChrZofV5. However, the presence of repeat units suggests that unplaced contigs `chrUn97886`, `chrUn83064`, `chrUn12635`, and `chrUn01845` and possibly `chrUn07087`, `chrUn06996`, and `chrUn06817` involve 5′-end telomeric junctions; and `chrUn10942`, `chrUn10872`, and `chrUn03315` and possibly `chrUn57207` involve 3′-end telomeric junctions.

Centromeres. From experience with difficult sequence and gaps from Phases 3/4, candidate loci for centromeres (or, more likely, pericentromeric repetitive sequences surrounding them) were known for several chromosomes. For an unbiased scan, a visual examination was made of the whole genome distribution of each common TRF canonical tandem repeat unit. Focusing on units tending to concentrate in at most one zone per chromosome, iterative examination of sequence in and near these zones (by dotplots, BLASTing, local reassembly, and visualization of genome-wide occurrences) led to an expanding collection of putatively centromere-associated sequences; these were consistent with candidate locations. The collection converged on "ChrZofCen" (given later), a single circular ~4 Kbp Type 1/Copia LTR retrotransposon with ~0.7 Kbp spacer, together with TRF canonical units `AAACATCTAG`, `AATCTGTGGTAGG`, `AAACATCTAGACACATCTAG`, and `AAACATCTAGACACATCTGG`, with some 5S rDNA sequence.

The plot below outlines chromosomes in gray; the *x*-axis is Mbp along '+' strands. Major assembly gaps (blocks of `N`s) are shaded light gray. Thicker segments outlined in black indicate the putative (peri)centromeric intervals given in the table on the next page. Red, orange, yellow, and green dots near top edges of chromosomes show TRF tandem repeats for canonical units `AAACATCTAG`, `AATCTGTGGTAGG`, `AAACATCTAGACACATCTAG`, and `AAACATCTAGACACATCTGG`, respectively. Purple dots show runs of consecutive hits to 19-mers of either strand of all IUPAC-ambiguity and curly brace expansions of ChrZofCen; vertical height within bounds of outlined chromosomes indicates number of consecutive hits, linearly from zero at the bottom of outlines to 1,594 (the maximum observed) just below the level at which green dots are shown.



Chr-	*Start ('+'*	*End ('+'*	*Nominal*	*Comments*
om.	*strand, bp)*	*strand, bp)*	*width (Kbp)*	
1	≈3,418,656	≈3,457,392	≈39	strong
2	2,093,247	2,141,774	49	strong, extra at 1,065,500–1,069,989
3	2,551,134	2,571,108	20	strong
4	2,648,641	2,651,949	3	possibly 2,719,643–2,722,347 (with asm. gap after) or 937,962–943,022
5	1,034,650	1,047,659	13	strong
6	709,495	716,341	7	strong, with assembly gap after
7	2,360,779	2,420,790	60	weak, with assembly gap inside
8	639,124	644,655	6	weak, with assembly gap after
9	in a gap	in a gap	?	no good candidates even though no large assembly gaps on this chrom.
10	860,629	862,963	2	weak
11	1,205,545	1,222,695	17	strong, with assembly gap inside
12	1,369,284	1,377,652	8	strong, with assembly gap inside
13	1,675,799	1,692,810	17	weak, with assembly gap inside; chromosome has large assembly gap
14	443,632	450,078	6	possibly 736,796–739,088 (with assembly gap before)
15	1,526,503	1,537,899	11	strong, with assembly gap before
16	490,261	510,147	20	strong, extra at 772,237–776,289
17	1,793,652 or 126,717	1,796,217 or 127,771	3 or 1	first option is at end of chrom., second with asm. gap after; 5 large asm. gaps
18	in a gap	in a gap	?	no good candidates; chromosome has three large assembly gaps
19	935,605	973,134	38	weak, with assembly gap inside

There are 39 unplaced contigs likely containing (peri)centromeric fragments: `chrUn{42003, 22516, 18154, 16591, 13058, 12366, 09183, 08437, 08040, 06312, 05306, 04914, 04275, 04018, 03492, 03384, 03059, 03028, 02729, 02724, 02655, 02649, 02593, 02484, 02398, 02352, 02284, 02246, 02034, 01939, 01933, 01883, 01678, 01641, 01499, 01429, 01415, 01238, 01183}`.

ChrZofCen (with IUPAC ambiguous nucleotides and '{*option*$_1$, *option*$_2$, …}' curly braces capturing the most common variations observed) consists of the following coding portion (which, in all expansions, starts and ends on a codon boundary with `ATG` and `TAA`):

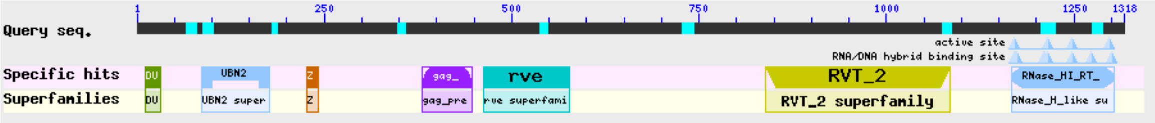
```
ATGACAGAACTGGAGAAGCTGGGTATCCCAArACTkAACGACCACAACTATGTCTTCTGGCACATCAAGATGCGAGCCTACCTyGTTGCAAGAGGAT
ACAGCGCAGCAATAACGAACGCAGAAGACGCCAACAGTGACAAGGCTCTTGCTTCCATCACTTTGGCTGTGGAAGATCATTTTCTACCTACAGTrTA
CAAwGCTGCAAGTGCGAAGGCAGCATGGGACGCGCTGGAGGCGTTGTTTCAGCAGCGGAGCGTTGCCAACCAGCTGAACCTCACGCAGGAACTGAAC
AACCTCACACTGCAGCCTGGGGAGACCATCACACAGCTACTTGCTCGTGCCAGAATCATATGGGAGCAGCTTAAGGCAGCTGGTATCGACAAGTCAG
AGCAGGAGGTGGCGTTATCAGTGTTGTCAGGACTTCCTGCCGACTTCAACACCTTAGTGACAGTACTACAGAATCAGTCTGGTCCmCTyACyCTGrG
TGGCATCCAGAAGGCTGTCTTGACAGAACAGCAACGTGCAAATAAGGTTGGGGCATCAACGTCTACTGCAGCAAGCACCAAGGCTTTCTACACTCAG
AACGGTCCCAACCrTGGCArGCTTGGTGACAGCGGTACCAGGACCAGCAACTT{,CAACCAGGGGAACrG}CAACACCAAGCAGCAGGAGCAGCGTA
AGTGCTACTACTGTGGCAAGAAGGGGCACCTGAAAAGGGACTGCAGAAAGAAGAAGGCAGACGAGCAGCGTGGCCCCAGTACCAAGGCTTCAACAAC
AATGGCATGGACTGCAGCCTGCAACACCAGCATCAGCCTCAGCTCAGGTACCTGGGTCCTCGACTCTGGAGCATCAAGACACGTCTGCAAAGAACGC
AGCCTGATGCAGAACCTGCAACAGCTGAACCAGCCAGTCTACATCACGTACGGCAACGGTAGCACAGGGGTGGCACAGACTATGGGGGAGGTTGTTC
TCAACGACAGGATCCGTCTACGGAACGTTTTGTTTGATCCCACTGCTGTTGGCAATCTCCTTTCCATCCsTACAGCAGCTGCryGTGGAGCACAGTT
```

```
TAACTTTGsAGCCArTTGCTGCACCATTCGAGTAAATGGCAGACTGGTGGCAATAGCACAGCAGCAwGAyGGTCACTAyTGCTTGCACTCTGAGCAw
rCAsAGTCAGCCACTGCACTGGCAGCCCAGACCCCGCAGCTGTGGCATCGTCGTTTTGGCCATCTCAGCTACCAGAATATGGCCAAGGTCCCCAACT
TGGTAACGGGCGTCCAAGTsCCAACTGArGCCTTTCAGGCAGCAGGTCAGCAGGTGTGTGAGCCATGTCTACTkGGCAAACAGACACGACTGTCTTT
CCCCGAGTCAGArACTGTCAGGCAGCAGyCACTkGArCTGGTGCATATGGACCTCTGTGGACCTCTyCCTGTCAAGTCACTTGGAGGCAGCCAGTAC
ATTGCTACGTTCCTGGAyGACTAyACAGGACTGTCAGTrGTGGCATTGCTCAAACAGAAGTCAGACATTTCyAArGTTGTGCCTGACGTCTTCAACA
TGCTAGAGAAACAGAGCAACAATCAGGTGAAGGGCGTCCGCACTGACAACGGCGGGGAGTATGTCAACAATGTGmTGAACAGCTACTACAGCAGCAA
GGGCATCATCGCACAGCACACAGTACCATACAGTCCTCAGCAGAATGGCAAGGCAGAAAGACTCAACCGAACCCTACTGGACAAGGCACGTTCCATG
CTGGCAGATGCArGGCTACCTTCTCAGCTrTGGGGTGAGGCCGTGGTAACAGCCAATTATCTTAGGAACCGTTCACCAGCAGCTGGCAAGACAGCAA
CACCCTGGGAACTGTTTTTTGGGTCACGGCCCTCTGTCTCTCATCTTCGCGTGTTTGGGGCCAAGGCGTTTGCACAGATCCCCAAGGAGAAACGTGG
CAAGCTGGACCCAAGGAGTCAGCGTGGCATCATGGTTGGATATGAGCCyAATGTAAAGGGGTACCGTCTACTGCTTCCAAACAACACCATCACAGTC
AGCCGGGACGTTGTATTTGATGAAGGTGACCAGCCAGGAGCArTAGACACCAACTTCTATCCAGACTTGGAAGATGAGCTTGATGTTACTGCAGCCA
TCAACACTGGATCTAATGCAGCACCTTCTGTCAATACTTCTGGAACAGCTGAGCCACCACCATCAGTTGCAGCACCCGTCGACCCACCAATTTCGGC
ACAGACCATGGAAAACGTGGGAGCCAGCAACAGCTCAACACCACAAGGCAGyGAGGAAGATCAGCATCAGCAATCACGTAGAAGTAGCCGGGCCAAC
ATTGGCATGGCACCAGGCAACTACTGGGAGGCCAACTACATTCCCACATCCAAGCGTACAGCTACCGGACTGTTGGCACAGACATCAGAAATTGTTG
AGCCAGCAACCTATGAmGAAGCACTACAGTCAGACTGTGCAGAGCAGTGGCAGCAAGCCATGGACAGCGAGTACGCATCGCTGATAGCCAATGGAAC
TTGGACCTTGGAAAAACCCCCAACAGACATTAGGCCCATCCCTGTCAAGTGGGTGTATAAGGTGAAACGTGACACCAGCGGGAACATTGAGCGGTTC
AAGGCACGCCTGGTGGCCAAGGGTTTTTGGCAACAGGAAGGTGTGGATTATGACGAAGTGTTCGCCCCGGTAAGCAAGTATGCTACCTTTCGGGCAC
TAATGGCCAAGGCAGCAGAAGAGGACATGGAACTACACAAATTGGATGTCAAGACTGCGTTCCTTCAAGGCAACCTGGAAGAAGATGTTTGGATCCA
GCAGCCTCGTGGCTACGAGGArGGCAGCAGTGAACTmGCCTGTCATCTwCAyAAACCTTTGTACGGGCTCAAGCAGGCyCCTCGrGCwTGGCATCAG
CGGCTACAACAGGAACTACTGGCAGTAGGCTACACAGCATCAGCAGCAGACCCCAGCCTGTACTGGTACTGCATCAACGGGGACTATGTGTACCTCC
TGGTCTAyGTGGATGATATCCTGATTGCAGCCAAGCAGCTTGAGTCAGTCAAGGCAGTCAAGCAGCAGCTrTTAGGCTTATTTGAGTCGCGTGACCT
TGGAGAAGCwACATCCTACCTTGGTATGAGCATTCAGCGCAACAGACAGACAGGCAyCATCAAGATyGGGCACCGACTCATGATCACAGAGTTACTG
GArArGTATGGyGCAGTmGACAGCAAAAThAAGTCArTACCACTGTCTCCATCTATCAArCTrGCyAAAGATGAAGGCGryCCCCTAGACAAGGAAC
ATTACCCTTACAGCCAACTGGTTGGGAGTCTCATGTACCTTGCAATCACCTCCAGGCCAGACCTCGCCTTTTCTGTGGGGGCTCTTGCACGCTACAT
GTCATGCCCAACCACwGTCCAyTGGCArGCAGCTAAGGGrGTrCTACGCTACTTGGGAGGAACCCTGGACTATGGCATCACCTTTGGTAGCGACAGC
AATGACCTCATTGGCTACTGTGACGCAGACTATGCGGGAGACACAGACACACGCAAGTCCACCAGTGGCTACATATTCATACTGCACGGAGGGGCCA
TyACkTGGAGTAGTAAGCGCCAGGCAACAGTTGCAGCmTCAACCACGGAGGCTGAGTACATGGCAGCAGCAGCAGCAGTCAAGGAAGCTCTATGGCT
GCGTACACTCTTGAGCGAGCTGCAGCTAGACATAGACAACATCACTATCATGGCAGACAACCAGTCAGCAATCAAGCTTCTGCGCAATCCTATCTCA
TCCATGAGAACCAAGCACATTGAyGTGGCTTATCACTTTGCTAGGGAACGCGTGGTGCGCAAGGAGGTTGTGTTCAGGTTCGTTTCCACAGAGAACA
TGGTGGCAGACATCATGACCAAGGCTCTGAGCGAAGTCAAGCATGTGCGATGTTGCAAGGGCATGGGGGTTGGAGTTTAA
```

followed by a more variable spacer region

```
AGAAACTTGAAATGCGTGGGAGCATCTTTGACAGTACATGCCTGACTGCGTGGGAGTGTTGAAATACGGCCTTTATTCAGTCAGACCTGCACTGCCA
GAATCCAGAAGTTGAGCATCTTTGACAGTACATGCCTGACTGCGTGGGAGTGTTGAAATACGGCCTTTATTCAGTCAGACCTGCACTGCCAGAATCC
AGAAGTTTCCAGATGGTTCTGGAAGyCCCCAGATGTTTCyAGATGTTTCyArATGTkTCy{,AAATGTGTCC}{,AGAAGTTTCTAGAGGTGTCTAG
ATGTTTCT}AGAATATTGGTGCATGACACGTGTCAGTCACTTTGTGGTr{,GTAGGAATCTGTG,GTAGGAATCTGTGGTAGGAATCTGTG}GTAGG
AATCTGTGGTAGGAATCTGTGGTAGGAATCTGTGGTAGGATTCCCAGTAGGTGAACACAGTTGCCAGTGGATTGCCATTGTGTCGTGAGTATATAAA
GACACAGACTTGTCCCAATCTGTAAyAyTGTCCAGCCCGAGysCCACCGAGGCCCCACGCTTAAACACAGACCGCAACACAGAGCTGAGGATACTGA
GTCGCTAGAACGACTwAGrCAACAGATTTCCATCAGGTTATGGGCCCACrCCCACACGCACAATCGCTGTGCTGCTCAGAAATTTGTTGTGTTCGGC
CATAAGTGTTGTGTACAGTTCGTCArCmAGGTCACd
```

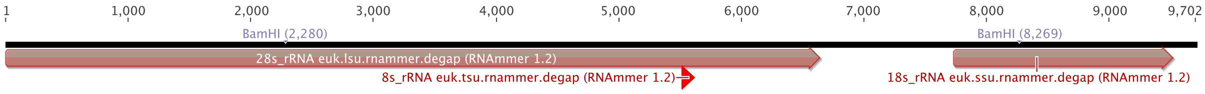
which then circles back to the beginning of the coding region. The Type 1/Copia LTR retrotransposal nature is clear from NCBI web conserved domain hits to its amino acid translation,



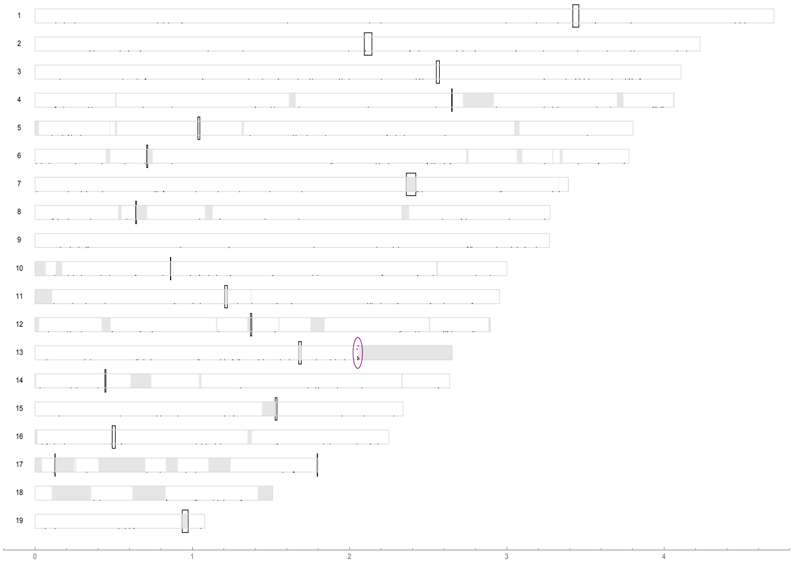
these being DUF4219 (a domain associated with the N-terminus of gag–pol proteins), UBN2 (gag of LTR Copia type), ZnF_C2HC/zf-CCHC (zinc knuckle associated with retroviral gag), gag_pre-integrs (part of gag lying just upstream of integrase), rve (the integrase core domain), RVT_2 (reverse transcriptase), and RNase_HI_RT_Ty1 (RNase H for Type I/Copia LTR retroelements), in that order. NCBI web BLASTX had best hit to filamentous green alga *Klebsormidium flaccidum* with second best organism being colonial green alga *Volvox carteri*.

Due to the difficulty of assembling such large-unit repetitive sequence occurring in multiple tandem arrays (and reads suggest each array consists of complex nested insertions of mixed orientation with some divergence), ChrZofV5 in (peri)centromeres — even when given as gapless pure A/C/G/Ts — may have considerable errors. Almost all putative (peri)centromeric intervals given in the table on the page before the last are associated with major assembly gaps and/or fine size differences between *in silico* BamHI fragment lengths and the optical map (with the assembly generally being too small; see also the discussion later about known assembly problems). However, borders and entry into pericentromeric sequences should be of quality comparable to the assembly generically, the optical map prevents massive errors (and constrains sizes), and sequence presently in ChrZofV5 should be representative. An estimate of the total size of (peri)centromeres was obtained in two ways. First, there is ≈195 Kbp of `N`-free sequence in the called intervals of the table and ≈231 Kbp of `N`-free sequence in the identified unplaced contigs/scaffolds (although all of such may not belong, as edges may be ordinary nuclear sequence), a total of ≈426 Kbp. Second, an analysis similar to that using Bowtie2 for the telomeric sequences (except selecting prepared Illumina reads as those having at least one 19-mer hit to either strand of all IUPAC-ambiguity and curly brace expansions of ChrZofCen or any rotation of (`AAACATCTAG`)$_2$, (`AATCTGTGGTAGG`)$_2$, `AAACATCTAGACACATCTAG`, or `AAACATCTAGACACATCTGG`) finds ≈290 Mnt of (peri)centromeric reads vs. the ~26.8 Gnt of nuclear reads, suggesting total centromeric nuclear sequence of ≈618 Kbp. Thus, *Chromochloris* may have a total of ≈0.5 Mbp of (peri)centromere, an average of ≈25 Kbp per chromosome.

Ribosomal DNA (rDNA). The canonical rDNA repeat unit for *Chromochloris* became apparent early in assembly during analysis of *k*-mers observed with high frequency, and is given as contig `chrRr`. It assembled as a 9,702 bp circular consensus which RNAmmer (48) annotates as follows (the consensus was oriented so that annotations fall on the '+' strand, and the de-circularizing linearization cut was placed just before the 28S rRNA annotation):


1
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
9,702
BamHI (2,280)
BamHI (8,269)
28s_rRNA euk.lsu.rnammer.degap (RNAmmer 1.2)
8s_rRNA euk.tsu.rnammer.degap (RNAmmer 1.2)
18s_rRNA euk.ssu.rnammer.degap (RNAmmer 1.2)

The plot below is similar to that shown earlier for centromeres, except there are no red, orange, yellow, or green dots formerly for various tandem repeats. Instead, purple shows length of consecutive hits to 19-mers of either strand of circular `chrRr`, with the top of chromosome outlines corresponding to maximum observed value 4,544.



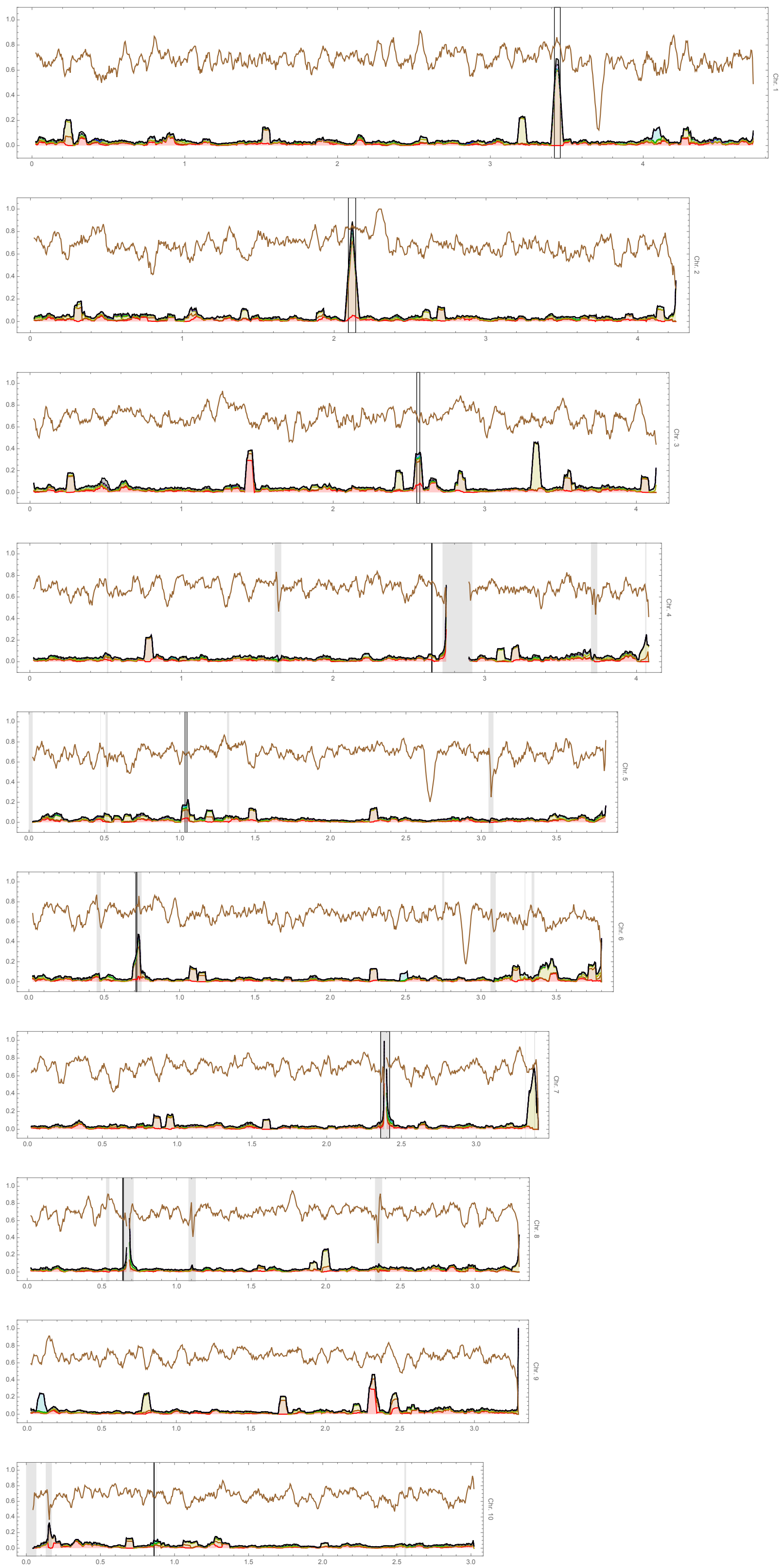
Typically, rDNA exists in at least one large tandem array; such sequence is, however, difficult to assemble. As with the (peri)centromeres, bordering sequence and entry into such regions is expected to be less problematic. From the plot, the tail of chromosome 13 (relative to its '+' strand) leads into a large assembly gap with rDNA sequence at the left border (indicated by purple oval). Further, the rDNA consensus contains two BamHI sites which, in circular form, produce fragments of sizes ~6.0 Kbp and ~3.7 Kbp that are both in the range in which the optical mapping worked well, and this region of chromosome 13 has an optical tandem repeat of ≈24× copies of an alternation of fragments ≈6 Kbp and ≈4 Kbp in size. This suggests that the large assembly gap at the end of chromosome 13, estimated to be ≈593 Kbp in size, begins with ≈24× copies of the rDNA unit (likely with divergence among copies); these copies would represent ≈233 Kbp or ≈40% of the gap. Various analyses (e.g., that of Table S1) assume this gap begins with 24× exact copies of `chrRr`.

Repetitive sequence. There are repetitive sequences beyond the telomeres, (peri)centromeres, and rDNA already discussed. The nuclear fraction of the ChrZofV5 assembly was analyzed with RepeatMasker 4.0.6open (using slow search and `gccalc` options with engine RMBlast+ 2.2.28) in combination with all of Repbase Update 2016-08-29 ("eukaryota") and *de novo* identified repeats from RepeatModeler 1.0.8open with RepeatScout 1.0.5, RECON 1.08, TRF 4.04, and RMBlast+ 2.2.28. About 6% of the assembly (excluding `N`-runs) was masked, mostly in interspersed repeats (~5.0% of sequence) as primarily LINEs (~2.0%), LTRs (~1.5%), unclassified elements (~1.2%), and DNA elements (~0.4%). The remainder was mostly simple repeats (~1.0%), with some satellites, low complexity sequence, and small RNA (total ~0.1%).

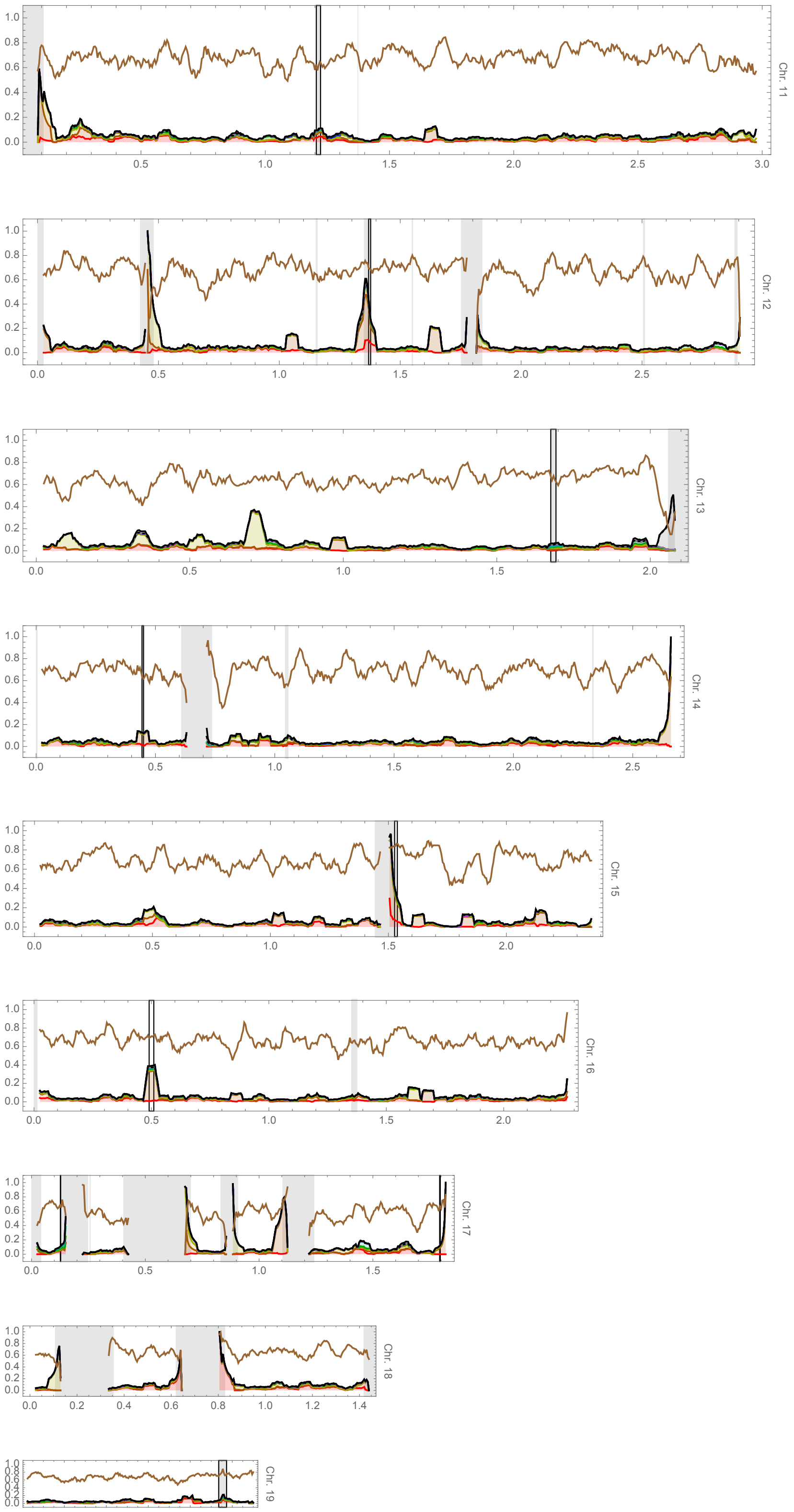
On the next two pages are per-chromosome plots of repeat and gene density with *x*-axes giving Mbp along '+' strands being divided into overlapping bins of 50 Kbp every 10 Kbp, and *y*-axes indicating fraction of non-`N` basepairs annotated in various ways. Repeats are shown stacked with bottom to top being LINEs (red), LTRs (orange), unclassified (yellow), simple (green), DNA elements (cyan), satellites (blue), RCs (magenta), low complexity sequence (gray), rRNA (dotted gray), tRNA (dotted black), and other/mixed class (black). Fraction of non-`N` basepairs annotated to the coding span of at least one ChrZofV5 gene model is given in brown. Major assembly gaps (large blocks of `N`s) are shaded light gray and vertical black lines indicate putative (peri)centromeres discussed earlier.

Gene density is rather uniform, and there are no grand scale gradients in genes or repeats as found in, e.g., *Arabidopsis*, a genome approximately twice as large that has megabasepairs of pericentromeric heterochromatin (49). Some smaller scale gradients in repeats are found near (peri)centromeres and especially large assembly gaps (e.g., the large gaps of chromosomes 17 and 18). There are a few localized concentrations of particular kinds of repeats. As apparent from Table S1, *Chromochloris*, like *Coccomyxa*, has relatively few repeats compared to other algal genomes of comparable size (*Chlorella* or *Monoraphidium*) and much fewer than larger genomes (*Chlamydomonas* or *Arabidopsis*).

Known assembly issues. From the hand word and detailed comparison of the final ChrZofV5 assembly to the optical map, 100 areas where the assembly has issues are known. (These are in addition to a likely number of very localized errors, e.g., individual basepairs; assembly polishing by variant detection using re-aligned reads is pending for the next assembly release.)



Repeat and gene density: chromosomes 1–10 (see description on previous page)



Repeat and gene density: chromosomes 11–19 (see description for previous page)

About half of issues (52/100) are represented in the ChrZofV5 assembly by runs of one or more N bases (typically with length a multiple of 1,000 bp sized approximately correct via the optical map); another half (47/100) are deviations in number of BamHI fragments or fragment lengths between the assembled sequence and optical map beyond norms; and a final one (1/100) is the optically-troubled tail of chromosome 5 mentioned earlier. Issues are detailed in Dataset S4 and summarized in Dataset S5. Below is a brief discussion of the summary and largest issues.

The largest assembly gaps (≥ ≈100 Kbp) are: one in the interior of chr. 4 (issue V5.04.4) of ≈193 Kbp; one at the beginning of chr. 11 (V5.11.1) of ≈107 Kbp; one at the end of chr. 13 (V5.13.6) of ≈593 Kbp of which, as already discussed, the first ≈220 Kbp of which is likely ≈24× copies of the rDNA repeat unit (so the amount missing is more like ≈373 Kbp); one in the interior of chr. 14 (V5.14.2) of ≈128 Kbp; three in the interior of chr. 17 (V5.17.2, .4, and .6) of ≈120, ≈295, and ≈138 Kbp; and two in the interior of chr. 18 (V5.18.1 and .2) of ≈248 and ≈208 Kbp. Chromosomes 17, 18, and 13 have by far the most assembly gap as a proportion of optical length. The presented sequence of chromosomes 1, 2, 3, and 9 is gapless, although, as discussed earlier under, e.g., centromeres, this does not imply they are perfect. Much of the "missing" sequence is expected to be among the unplaced contigs/scaffolds.

OpGen did not observe the right end of chromosome 5 (issue V5.05.10); its map ends in an inverted optical repeat of ≈150 Kbp per arm. From patterns of chromosomal coverage by reads and detailed hand examination of Illumina and PacBio reads aligning and partially aligning in this region, it was determined after the ChrZofV5 assembly was frozen and this publication was initially submitted that the likely resolution is this inverted repeat is much larger — fully ~564 Kbp per arm — with the right arm exiting directly into telomere repeats. Using 1-based inclusive-inclusive '+' strand coordinates in ChrZofV5, a full left arm is given as chr05:3230407–3794116 and a full spacer between the two arms is given as chr05:3794117–3795139, but the end chr05:3795140–3801251 only gives ~6 Kbp of a right arm (and with chr05:3796510–3798042 being naked single-read PacBio sequence). A quick patch is to tack revComp(chr05:3230407–3788934) :: AAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGG GTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGG onto the end of ChrZofV5 chr05, making chr05 longer by 558,602 bp, but it is planned for the next genome release to use the PacBio reads to phase the two arms (which appear to have some variation) and give a better representation (as the sequence currently in ChrZofV5 is presumably randomly phased). Note there are 155 current gene models affected (Cz05g32080, …, Cz05g37220 in the left arm and Cz05g37230 and Cz05g37240 in the right); there will be another ~153 once the rest of the right arm is added. Except where explicit, this publication assumes a genome with most of the right arm absent.

ChrZofV5 chromosomes 1 to 19 total 57,719,290 bp (including N placeholders); the optical map totals 57,763,775 bp (with only the first half of the chromosome 5 optical repeat counted). These agree to ≈1 part per thousand and, when quoting lengths as fractions of the nuclear genome, it does not matter much which is taken as reference whole. (The total differs by under 45 Kbp and single chromosomes by under ±42 Kbp.) About 5% of total (≈3 Mbp) is missing over the 52 runs of Ns; the unplaced contigs/scaffolds presumably provide ≈2.4 Mbp of this (≈80%). Over the 47 BamHI fragment disagreement issues, assembled sequences are estimated to be missing ≈512 Kbp and have ≈45 Kbp extra; this is under 1% of total and a smaller class of problem than the runs of Ns. Thus, ≈6% of total is missing or otherwise troubled, but ≈94%

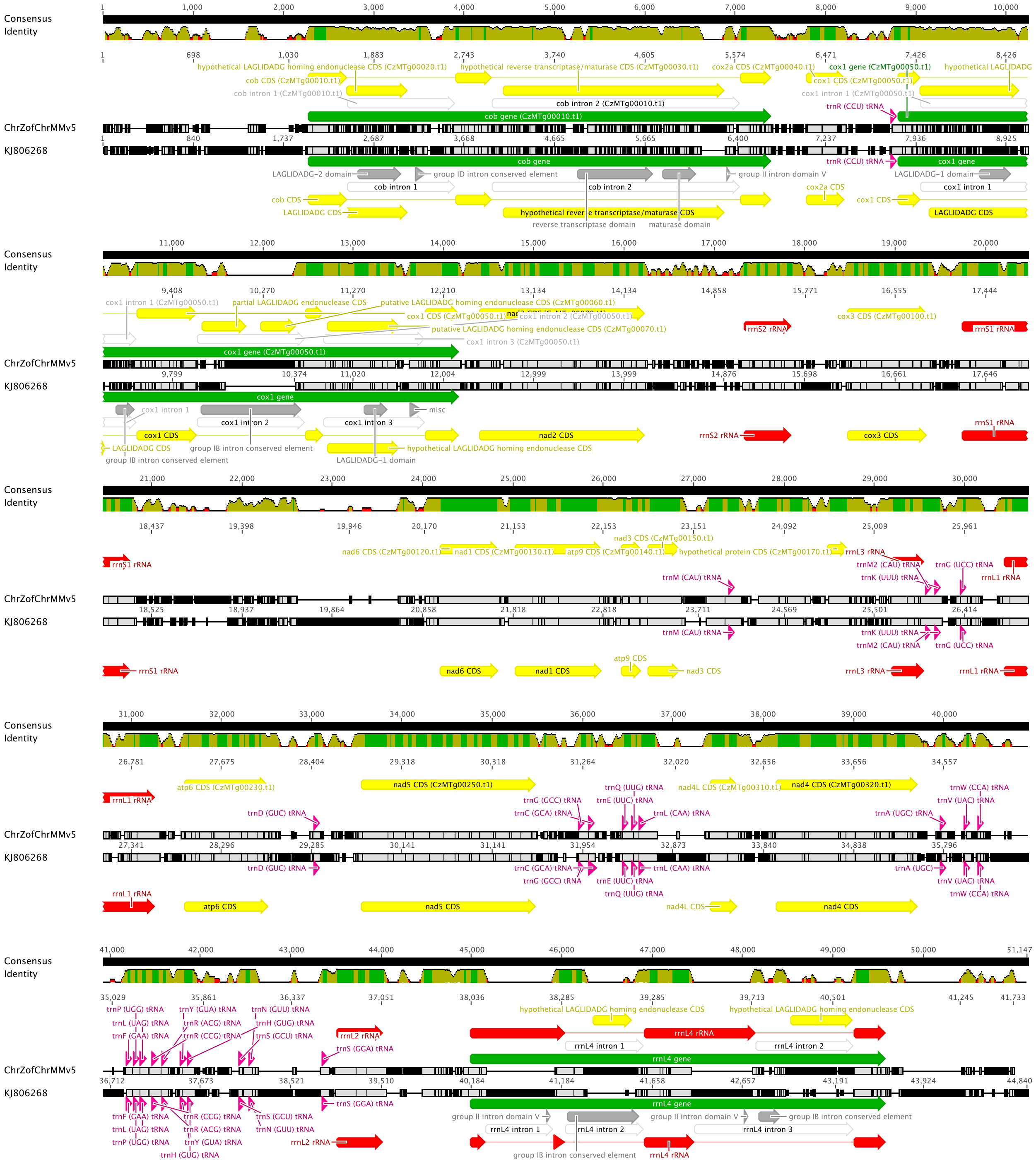
is placed and in tight agreement with the optical map. Although current data is not exhausted and additional refinements can be made, in the interest of timely availability to the community of the already high quality genome, the ChrZofV5 version is being publically released.

Genomes of the chloroplast and mitochondrion

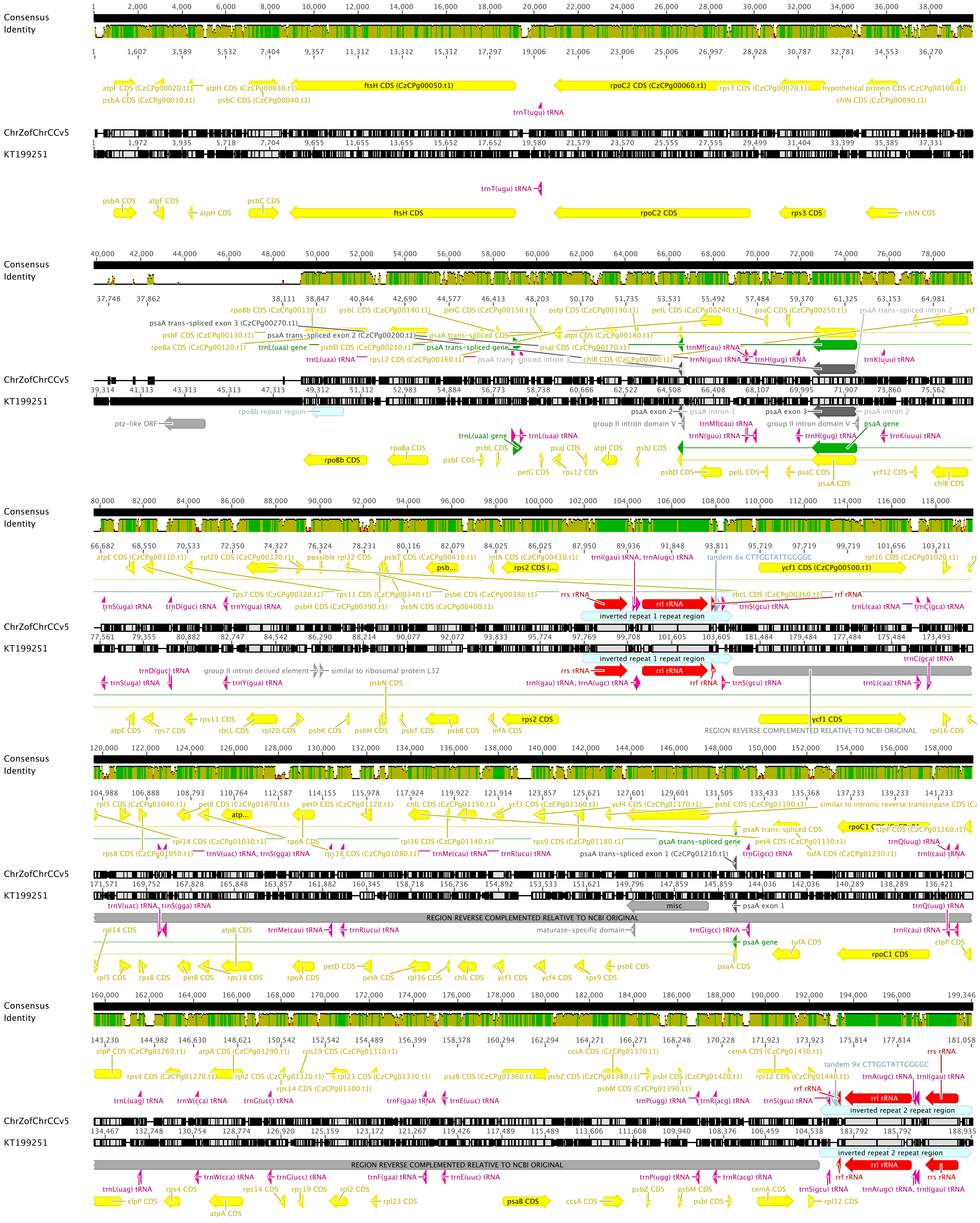
Assemblies of organelles took place between nuclear assembly phases and also required multiple hand-managed passes (as they were not assembled whole by any of the automatic processes). Various methods were used to identify potentially relevant contigs and reads, including relatively high coverage, low G+C content, alignments and synteny to existing NCBI chloroplast and mitochondrion sequences, and alignments to seed contigs once some were in hand.

Mitochondrion. The mitochondrion (for SAG 211-14, the strain of this study) was completely assembled as a single circular 41,733 bp contig `chrMt` with no IUPAC ambiguous nucleotides; the strand orientation and linearizing cut were chosen to agree with NCBI accession KJ806268.1, the 44,840 bp complete mitochondrion of *Chromochloris zofingiensis* strain UTEX 56. Annotation of protein-coding genes, tRNAs, and rRNAs of `chrMt` was by BLASTN/BLASTX and BLASTP to the NCBI 'nt' and 'nr' databases, tRNAscan-SE, RNAmmer, Rfam, syntenic alignments to closely related known sequences (e.g., to KJ806268.1), and visual examination of RNA-Seq alignments (which suggest some UTRs, although these were not kept in the final annotations). The overall structure of `chrMt` is highly similar to KJ806268.1, having the same major protein-coding genes, tRNAs, and rRNAs in the same order, however there is considerable divergence at the nucleotide level with a global pairwise alignment (Geneious 93% similarity cost matrix, gap open penalty 30, gap extension penalty 1; see figure on next page) only ~66% identical. Divergence is concentrated intergenically and the splicing structure of rrnL4 is different. Globally aligning just the coding sequences results in ~98% nucleotide identity. Translating the coding sequences via NCBI genetic code #22 (the *Scenedesmus obliquus* Mitochondrial Code) and globally aligning (Geneious BLOSUM62, gap open penalty 12, gap extension penalty 3) estimates ~99% amino acid identity.

Chloroplast. Similarly, the chloroplast (for strain SAG 211-14) was completely assembled as a single circular 181,058 bp contig `chrCp`, also with no IUPAC ambiguous nucleotides; the strand orientation and linearizing cut were chosen to be in agreement with NCBI accession KT199251.1, the 188,935 bp complete chloroplast of *C. zofingiensis* strain UTEX 56. Again, annotation of protein-coding genes, tRNAs, and rRNAs was by BLASTN/BLASTX and BLASTP to NCBI 'nt'/'nr', tRNAscan-SE, RNAmmer, Rfam, syntenic alignments to closely related known sequences, and visual examination of RNA-Seq alignments (which again suggested some UTRs, although as before these were not kept in the final annotations). As with many chloroplast genomes, there is a large rRNA-related inverted repeat (~6.7 Kbp in SAG 211-14, ~6.4 Kbp in UTEX 56) separating two single copy regions. It is difficult to resolve the arms with short reads; they assembled as identical except for a tandem repeat `CTTGGTATTGGGGC` estimated as 8× in the first arm and 9× in the second (where SAG 211-14 inserts ≈300 bp relative to UTEX 56). The relative strand orientation of the single copy regions is ambiguous, and no PacBio reads were found able to resolve this. The single copy regions were assembled in opposite relative strand orientation compared to KT199251.1, and so in further comparisons the second single copy region of KT199251.1 was reverse complemented.



Global nucleotide alignment of mitochondrion genomes `chrMt` *(Chromochloris strain SAG 211-14) and NCBI KJ806268.1 (Chromochloris strain UTEX 56); see discussion on previous page*



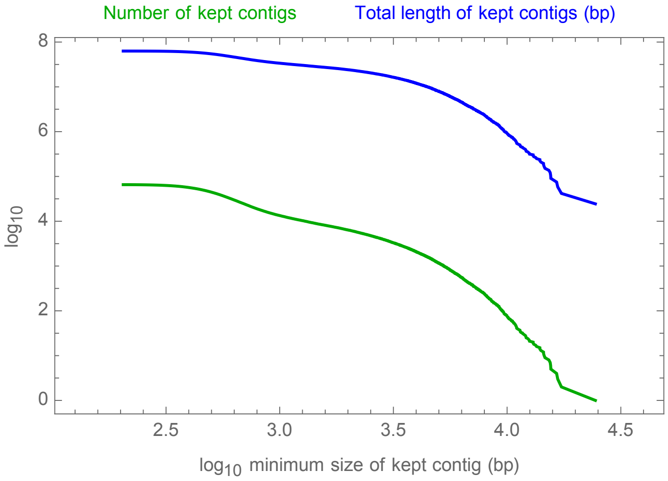
Global nucleotide alignment of chloroplast genomes `chrCp` *(Chromochloris strain SAG 211-14) and NCBI KT199251.1 (Chromochloris strain UTEX 56); see discussion on next page*

With the second single copy region of KT199251.1 reverse complemented, the overall structure of `chrCp` is highly similar to KT199251.1, with the major protein-coding genes, tRNAs, and rRNAs again in the same order. Aligning in the same way as with the mitochondrial genomes, global alignment gives overall nucleotide identity of ~83% and global alignment after restriction to coding sequences gives ~98%; divergence is again concentrated intergenically. The largest difference is the loss in SAG 211-14 of almost all of a ~9.3 Kbp region in UTEX 56 annotated as containing a ptz-like ORF. Translating the coding sequences via NCBI genetic code #11 (the Bacterial, Archaeal, and Plant Plastid Code) and globally aligning results in ~97% amino acid identity, with lower percent identity in the larger genes (e.g., ftsH, rpoC2, and ycf1). The gene psaA remains trans-spliced (with RNA-Seq in concurrence); an *in silico* effort to identify a homolog of the *Chlamydomonas* tscA gene involved in this process was unsuccessful.

From the Bowtie2-based analysis introduced in the telomere discussion earlier, coverage on `chrMt` and `chrCp` from prepared Illumina reads is ≈1,280× (≈0.2% of sequencing effort) and ≈890× (≈0.6%), respectively. Coverage of PhiX is ≈150,000× (≈2.8%), and the nuclear genome (chromosomes, rDNA, and unplaceds) is ≈460× on average (≈93.9%). The remaining ≈2.6% of effort is in reads that did not align; ≈1.5% is accounted for in a re-alignment to Illumina inline controls, leaving ≈1.1% of effort unaligned. At nuclear coverage, this could be ≈0.7 Mbp of additional sequence, very close to the ≈0.6 Mbp more expected beyond current unplaced contigs/scaffolds as discussed under “Known assembly issues” above. The high fraction (≈98.9%) of original reads accounted for is encouraging; there is not much sequence missing from the ChrZofV5 assembly, even if it is not yet in perfectly contiguous form.

Transcriptome assembly used in training AUGUSTUS

To assist with training the AUGUSTUS *ab initio* gene modeler for *Chromochloris*, a draft transcriptome was *de novo* assembled from the 151+151 nt pool of ~397M read pairs from the fourteen RNA-Seq sub-libraries described earlier, using the Ray assembler with k = 51. Such *de novo* transcriptome contigs are generally presented in random strand and codon frame, generally contain UTRs, and may contain introns (and many of large number of shorter, lower-coverage contigs may be wholly introns, and introns may change codon frame). To bootstrap AUGUSTUS, PASA 2.0.2 was used to extract a training set of genes (50).



Details of Table S1

Nuclear genomes. Sequences and annotations (especially those of model organisms) are often updated after initial publication, and details of definitions and statistical analyses can often greatly affect summaries. For these reasons, Table S1 was completed by analyzing freshly-downloaded current copies of reference genome sequences and gene models and uniformly applying the same criteria and methods to every organism rather than, e.g., copying nominal quantities from existing publications. Sources of nuclear genomes and annotations were TAIR10 for *Arabidopsis thaliana* ("*AraTha*"), JGI Phytozome 5.5 for *Chlamydomonas reinhardtii* ("*ChlRei*"), ChrZofV5 of this work (with 24× copies of the rDNA unit) for *Chromochloris zofingiensis* ("*ChrZof*"), JGI Phytozome 2.0 for *Coccomyxa subellipsoidea* C-169 ("*CocSub*"), JGI release 2014-08-18 with 'best genes' for *Chlorella* sp. NC64A ("*Chlore*"), and NCBI accessions KK100223.1–KK106940.1 for *Monoraphidium neglectum* ("*MonNeg*").

Sequenced genome size: number of non-`N`/`n` bases in assembly (other IUPAC ambiguities were retained), rounded to nearest Mbp. *Sequenced genome presentation:* a "scaffold" is defined as a nucleotide sequence having at least one `N`/`n` (with other IUPAC ambiguities being irrelevant) and presuming every other sequence to be a "contig". For *CocSub*, all sequences are called "scaffolds" in distributed files and chromosome vs. arm vs. unplaced is not indicated; however, the distinctions are clear from presence of telomere-associated repeats at one, both, or neither sequence edge, and the number of chromosomes plus half the number of arms as thus determined equals the stated 20 chromosomes in the associated genome paper (51), which also mentions that, via Southerns, the pairing of half the arms was determined. *Genome project primary initial strategy, average basepair coverage at earliest stage:* per best evidence available and literature, including *CocSub* (51), *AraTha* (49), *ChlRei* (52), *Chlore* (53), and *MonNeg* (54).

Scaffold N50 (taking genome size as sum of scaffolds as-are): ordering scaffolds by decreasing size (and keeping all IUPAC ambiguous nucleotides), take scaffolds until total size is at least as large as half total size of all scaffolds, and report size of the smallest taken scaffold after rounding to the nearest Kbp. This was not performed for assemblies at chromosome/arm scale, as this quantity is then essentially as large as it can be and is controlled by the organism's distribution of chromosome sizes and is no longer connected to assembly quality. *Contig N50 (taking genome size as sum of contigs as-are):* form "contigs" by splitting scaffolds at every `N`/`n` (tolerating other IUPAC ambiguities) and removing all `N`/`n`s; order contigs by decreasing size, take contigs until total size is at least as large as half total size of all contigs, and report size of the smallest taken contig after rounding to the nearest Kbp. *Number of chromosomes:* per best evidence available and literature. For *Chlore*, although not mentioned in the associated genome paper (53), the largest scaffolds are large and one can look for telomere-associated repeats; 11 of their scaffolds begin with such a repeat and 7 end with one (and none have both ends thus associated)… this is more or less consistent with the genome paper's determination of 12 chromosomes by Pulsed-Field Gel Electrophoresis (PFGE), with chr. 12 being difficult.

Protein-coding genes are taken as those directly declared as such by the annotations; in cases (*MonNeg*) without a direct indication, a GFF file `gene` was taken as protein-coding if and only if it had non-empty intersection with at least one GFF file `CDS` interval. Three of the releases here (for *ChrZof*, *MonNeg*, and *CocSub*) do not provide multiple transcript models per gene locus. (Although the *CocSub* release includes versions of files named so as to distinguish all models vs. "primary transcript only", such versions are the same and no protein-coding locus is actually modeled with multiple isoforms.) For *AraTha*, when desired, the canonical transcript model for each gene locus is per TAIR's file `TAIR10_representative_gene_models.gz`. For *ChlRei*, when desired, the canonical model is that marked `longest=1` in the annotation GFF files (and all GFF files in the release agree on this designation). For *Chlore*, there is no issue since the only gene models used in this work are those from the release's 'best genes' files. Note that *MonNeg* is a highly fragmented assembly and one may expect (in agreement with the BUSCO analysis of main text Fig. 3A) its gene models to suffer due to, e.g., true coding sequences often reaching edges of assembly sequences; for this reason, numerous of its gene-related summary statistics may be more divergent from "truth" than for the other organisms.

Percent G+C in sequenced genome: using only `A`/`C`/`G`/`T` nucleotides (no IUPAC ambiguities) in an all-uppercase version of the assembly, report fraction (# `C`+`G`) / (# `A`+`C`+`G`+`T`) as a percent, rounded to the nearest integer. *Basepairs called as coding (in any transcript model) in sequenced genome:* over all transcript models of all protein-coding genes, take union of coding sequence bases (ignoring strands) to get a target subset of assembly basepairs; restrict to the `N`/`n`-free fraction of this subset and the whole assembly (other IUPAC ambiguities being tolerated), and report percentage of the whole in the subset after rounding to the nearest integer. *Percent G+C in basepairs called as coding (in any transcript model):* over all transcript models of all protein-coding genes, take union of coding sequence bases (ignoring strands) to get a target subset of assembly basepairs; using only `A`/`C`/`G`/`T` bases (no IUPAC ambiguities) in an all-uppercase version, report fraction (# `C`+`G`) / (# `A`+`C`+`G`+`T`) as a percent, rounded to the nearest integer. *Number of "complete" called protein-coding gene loci (collapsing transcripts):* same as the row "Number of called protein-coding gene loci (collapsing transcript forms)" except restricted to coding sequences that satisfy all of the following: (1) are pure `A`/`C`/`G`/`T` (i.e., contain no IUPAC ambiguities); (2) begin and end on codon boundaries; (3) start with `ATG`; (4) end with `TAA`/`TAG`/`TGA`; and (5) do not contain an internal `TAA`/`TAG`/`TGA` codon.

Number of rDNA units estimated to exist in true monoploid genome: MonNeg and *AraTha* are via (54) and (55); for *ChrZof*, this work as, e.g., already described in subsection *Ribosomal DNA (rDNA)*. For *ChlRei*, the original genome paper (52) contains some information but not quantitation. Seven paired-end 76+76 nt Illumina GA-II lanes of a *Chlamydomonas* genomic library were available from an unrelated project. Extremely high coverage 39-mers from the reads were *de novo* assembled, rDNA-related seed contigs selected via NCBI web BLASTN, paired-end reads having at least one 31-mer from the seed contigs and seen multiple times were extracted and re-*de novo* assembled to obtain a 6,543 bp consensus chunk of a presumed *Chlamydomonas* rDNA unit. The chunk contains a whole 18S followed by a whole 28S. Comparison of median Jellyfish 39-mer coverages for the consensus chunk of rDNA unit vs. some generic "1×" ordinary sequence in the nuclear genome (that `N`/`n`-free chunk of chr. 1 with length 440,320 bp, with a coverage threshold to remove empirically non-unique regions) provides an estimate of rDNA unit copy number as 840× (independent of the chunk's tandem circle not being closed), and ~5.5 Mbp as a lower bound on total length (dependent on the fraction of the unit the chunk represents).

Number of tRNAs called in sequenced genome: counts are for all types (including with introns, unclassified, selenocysteine, and pseudo). For *AraTha*, the TAIR release contains explicit tRNA annotations, and the table entry '631' counts these. For the other organisms, even though, e.g., the original genome papers generally discuss tRNAs (implying that predictions were made), the annotation releases do not identify tRNAs and so for this work *ab initio* scans with tRNAscan-SE 1.3.1 were performed with default parameters. (This scan finds 639 for *AraTha*.) For *MonNeg*, the *ab initio* scan found 38 in the nuclear genome, 29 in the chloroplast, and 23 in the mitochondrion, while the original genome paper (54) states "40 + 1× Pseudo Ser-tRNA" for nuclear, "29 + 1× Pseudo Leu-tRNA" for the chloroplast, and "21 + 1× Pseudo Met-tRNA" for the mitochondrion in its Table 3 but shows 23 in its Figure 5. For *ChlRei*, the '259' shown in Table S1 is taken from the original genome paper (52), as even though the current JGI 5.5 release does not contain tRNA annotations, the original genome paper states that tRNAscan-SE is known to overestimate in *Chlamydomonas* due to tRNA-associated SINE retrotransposon elements; the *ab initio* scan predicts 353 tRNAs in the current nuclear assembly. Regarding *ChrZof*, the scan only identifies 75 tRNAs (Dataset S8) — more than *Chlore* and *MonNeg*, the small algal genomes of high G+C and moderate repeat content, and similar to *CocSub*, the other algal genome of of moderate G+C and low repeat content, but much less than the relatively large genomes of *AraTha* and *ChlRei*; there are no large clusters, although there are runs of up to four on the same chromosome with spacing smaller than would be expected at random (e.g., with some adjacencies closer than 1 Kbp).

Organism:	# selenocysteine:	# pseudo:	# undetermined:	missing std. AAs:
CocSub	0	3	4	none
ChrZof	0	0	0	none
AraTha (ab initio)	0	8	1	none
ChlRei (ab initio)	1	2	2	none
Chlore	1	0	0	Ile
MonNeg	2	1	1	Asn, Glu, Trp, Tyr

From the *ab initio* scans, all standard amino acids are covered for all six organisms, except for one in *Chlore* and four in *MonNeg*, perhaps because these are the most fragmented assemblies, or perhaps due to tRNAscan-SE misclassifications as selenocysteine/pseudo/undetermined (of which there are exactly one and four in *Chlore* and *MonNeg*, respectively). The phylogenetic profile of anticodons (ignoring predicted pseudogene status) is as follows: universal in all six = AGC, AGG, AGT, CAA, CAC, CAT, CGC, CTG, CTT, GAA, GCA, GCC, GTC, GTG, TCG; missing from all six = AAA, ACA, ACT, ATA, ATG, ATT, CTA, GAC, GCG, GGC, GGG, GGT, TTA; missing from just *MonNeg* = AAC, AAG, ACG, CAG, CCA, CGA, CGG, CTC, GTA, GTT, TAA, TGG; missing from *MonNeg* and *Chlore* = AAT, CCG, CCT, TAC, TCC, TGA, TGC, TGT, TTC, TTG; missing from just *Chlore* = CCC, CGT, GCT, TAG, TAT, TCT; missing from *MonNeg*, *Chlore*, and *ChrZof* = AGA, TTT; just *AraTha* = ACC, GAG; just *ChlRei* = ATC; just *MonNeg* = GAT; just *AraTha* and *CocSub* = GGA; and just *ChlRei*, *Chlore*, and *MonNeg* = TCA.

Number of amino acids: {average, median}: gene models are taken without question (e.g., even if one does not start with a start codon, end with a stop codon, has coding sequence not a multiple of three nucleotides in length, the coding sequence contains IUPAC ambiguities, the coding sequence is very long, …) and the result rounded to the nearest integer. *Number of* *exons containing coding sequence: {average, median}:* gene models are taken without question (e.g., no matter how many exons they have) and the result rounded to the nearest tenth. *Exon length (restricted to coding sequence): {average, median}; Intron length (between exons with coding sequence): {average, median}; Percentage with at least one intron (between exons with coding sequence):* same comments as for "Number of amino acids: average".

% of seq. basepairs RepeatMasker'd with {Repbase Update "eukaryotic", RepeatModeler, RepeatModeler + Repbase Update "eukaryotic"}: the RepeatMasker/Repbase/RepeatModeler analysis discussed earlier for *ChrZof* under subsection *Repetitive sequence* was applied to the other five organisms with the same parameters. Masking was variously with just known repeats (Repbase only), just *de novo* repeats from RepeatModeler, and the combination of the two.

Chloroplasts. For chloroplast genomes, reference sequences and annotations were as follows. *CocSub:* NCBI accession NC_015084.1 (with one sequence gap and lacking a large inverted repeat) and annotations. *ChrZof:* `chrCp` of the ChrZofV5 release of the present work. *AraTha:* NCBI accession AP000423.1 sequence and annotations. *ChlRei:* NCBI accession FJ423446.1 sequence and annotations. *Chlore:* NCBI accession KP271969.1 sequence (lacking a large inverted repeat) and annotations. *MonNeg:* NCBI accession CM002678.1 sequence, but with annotations from Fig. 4 of the original genome paper (54) as the annotations deposited at NCBI are manifestly highly incomplete. *Sequenced genome size.* The number of non-`N`/`n` bases in the assembly (other IUPAC ambiguities being tolerated) is reported, rounded to the nearest Kbp. *Number of annotated protein-coding genes, including hypotheticals; Number of annotated {rRNAs, tRNAs}:* if the genome contains large repeats (as common in chloroplasts), genes are counted as +1 copy for each copy of the parent repeat. For tRNAs, as with the nuclear genome, if no annotations were provided, an *ab initio* tRNAscan-SE scan was performed (and all types counted). *Percent G+C in sequenced genome:* using only `A`/`C`/`G`/`T` bases (no IUPAC ambiguities) in an all-uppercase version of the assembly, the fraction (# `C`+`G`) / (# `A`+`C`+`G`+`T`) is reported as a percentage rounded to the nearest integer.

Mitochondria. For mitochondrial genomes, reference sequences and annotations were as follows. *CocSub:* NCBI accession NC_015316.1 sequence and annotations. *ChrZof:* `chrMt` of the ChrZofV5 release of the present work. *AraTha:* NCBI accession JF729201.1 sequence and annotations. *ChlRei:* NCBI accession NC_001638.1 sequence and annotations. *Chlore:* NCBI accession NC_025413.1 sequence and annotations. *MonNeg:* NCBI accession CM002677.1 (with two sequence gaps) and annotations. Rows are the same as for chloroplasts, except for the following note not already mentioned elsewhere: the *MonNeg* mitochondrial sequence has no rRNA annotations (and two sequence gaps); RNAmmer does not find any rDNA, but Rfam finds four zones with LSU/SSU fragments.

Calling of protein-coding gene families across the six organisms of Table S1

To call gene families (Datasets S9–S18) simultaneously across *AraTha*, *ChlRei*, *ChrZof*, *CocSub*, *Chlore*, and *MonNeg* (the six organisms of Table S1), the amino acid sequences of the genes corresponding to row "Number of called protein-coding gene loci (collapsing transcript forms)" of Table S1 were collected. Alignment seeds were formed by running NCBI BLASTP+ 2.4.0 with E-value threshold 10^{-5} and soft masking (segmasker window 12, locut 2.2, hicut 2.5) on both queries and subjects (and otherwise defaults, including BLOSUM62 scoring). For every distinct ordered pair (*query*, *subject*) with at least one BLASTP+ result, global Needleman–Wunsch alignment was performed with the C++ library Parasail (BLOSUM62 scoring with gap open and extend penalties 10 and 1, respectively). Compared to the local alignments of BLASTP, the global alignment score captures not only sequence similarity, but also aspects of the fraction of the entirety of query and subject aligned and the ordering of homologous fragments (e.g., component protein domains).

In the first phase, “self-prefamilies” were formed within each organism. For Parasail-aligned pairs of genes (*query*, *subject*=*query*) with global alignment score $s \geq 16$, keep as “tentative arcs” those Parasail pairs (*query*, *subject in same organism except query itself*) with global alignment score ≥ 85% of *s*. Remove tentative arcs (*query*, *subject*) for which (*subject*, *query*) is not a tentative arc, so as to obtain unordered pairs {*gene*, *different gene in same organism*} that constitute edges in an undirected graph. Partition vertices of this graph (the pieces of this partition being the self-prefamilies) by subdividing the vertices of each connected component as follows: (1) find all maximal cliques in the connected component; (2) keep only cliques of maximum size; (3) expand each clique to also contain those vertices in the connected component that are adjacent to at least half the vertices in the clique; (4) keep only expanded cliques of maximum size by number of vertices in them; (5) group vertices in the union of the surviving expanded cliques by their combination of membership status in the surviving expanded cliques, these groups becoming pieces of the final partition; and (6) recurse [going back to (1)] on any vertices remaining. Finally, each gene in the organism not represented is added as a singleton self-prefamily (of size 1). Self-prefamilies involve 1 to 31 genes (but only 1 to 4 genes each when restricting to sizes occuring ≥ 10 times in any single organism, and only 1 or 2 genes each when restricting to sizes seen ≥ 100 times in any single organism). The percent of genes in self-prefamilies of size ≥ 2 is ~8.4% and ~4.6% in the large genomes *AraTha* and *ChlRei*, respectively; ~2.9% and ~2.2% in the algal moderate G+C content genomes *CocSub* and *ChrZof* of low repetitive sequence fraction, respectively; and ~1.1% and ~0.9% in the algal high G+C genomes *Chlore* and *MonNeg* of moderate repetitive sequence fraction, respectively.

Self-prefamilies exhibit evidence of tandem duplication events in all six genomes. For example, consider self-prefamilies of size exactly 2. (Across organisms, this is ~73% to ~96% of self-prefamilies of size ≥ 2.) Given such a self-prefamily, classify it as type “Far” if the two genes are on different sequences in the reference genome or the midpoint of the bounds of their coding sequences are ≥ 20 Kbp apart; otherwise, classify it as type “Near$^+$” if the two genes are on the same strand or “Near$^-$” if they are on opposite strands. There is enrichment for Near$^-$ and larger enrichment for Near$^+$ in every organism:

Organism:	# observed:			random expectation:			observed / expected:		
	Far	Near$^+$	Near$^-$	Far	Near$^+$	Near$^-$	Far	Near$^+$	Near$^-$
AraTha	714	214	30	~958	~0.181	~0.170	~0.75	~1,184	~176
ChlRei	146	58	30	234	0.040	0.041	0.62	1,454	725
Chlore	20	16	6	42	0.017	0.018	0.48	943	328
ChrZof	104	12	8	124	0.040	0.040	0.84	298	199
CocSub	41	13	1	55	0.020	0.023	0.75	649	44
MonNeg	59	9	1	69	0.006	0.006	0.86	1,512	174

In the second phase, “prefamilies” are formed — these target orthologs (“primaries”) and generally involve more than one organism. For Parasail-aligned pairs (*query*, *subject in different organism*) sharing the same query, drop all these pairs if the best global alignment score s is < 16 and otherwise keep only pairs with global alignment score ≥ 97% of s. Replace kept ordered pairs of genes (*query*, *subject*) with ordered pairs (*self-prefamily of query*, *self-prefamily of subject*) and thin ordered pairs seen more than once down to a single copy. Taking these as the new “tentative arcs”, follow the same procedure as used to form self-prefamilies, except the resulting partition pieces now constitute the prefamilies. Each of these involves 1 to 15 self-prefamilies; ~67% and ~90% of genes in multi-organism prefamilies belong to prefamilies with at most 1 and at most 2, respectively, self-prefamilies per organism.

In the third phase, final families are formed — with paralogs now also targeted as “additional” genes in each family — by merging into each multi-organism prefamily zero or more single-organism prefamilies. Each single-organism prefamily S is considered independently one at a time: for each gene a in S, gather Parasail alignments (a, *gene b in a multi-organism prefamily*) and (*gene b in a multi-organism prefamily*, a), keep only alignments with maximum global aligment score, and note the multi-organism prefamilies that surviving b belong to; if exactly one multi-organism prefamily M is noted after all a are considered and at least one kept alignment was seen with strictly positive global alignment score, then S is merged into M as additional genes (and otherwise S is left alone). 5,258 multi-organism prefamilies receive merges, each 1 to 196 times, with ~88% of these ≤ 6 times. There are 41,328 final families (these partitioning all 27,206 + 17,741 + 15,344 + 9,629 + 9,791 + 16,734 = 96,445 genes from *AraTha*, *ChlRei*, *ChrZof*, *CocSub*, *Chlore*, and *MonNeg*, with each gene belonging to exactly one final family), with 30,838 and 10,490 involving single vs. multiple organisms, respectively. Of the 10,490, 5,012 have ≤ 1 gene (primary + additional) per organism and 7,904 have ≤ 2 genes. The largest families are of various histone proteins.

Reference genome:	% of reference genes that belong to multi-organism families:	Same, except multi-orgo. family restricted to having ≤ 2 genes (primary+add’l) per organism:	Same, except multi-organism family has ≤ 1 gene per organism:
AraTha	~60%	~18%	~7%
ChlRei	64%	37%	21%
Chlore	83%	52%	29%
ChrZof	73%	47%	27%
CocSub	77%	50%	26%
MonNeg	63%	39%	20%

Details of main text Fig. 3B/C

Phylogram. The 813 protein-coding gene families (called across the six organisms of Table S1) that have no additional genes and exactly one primary gene in each of the six organisms were identified. Because of the highly fragmentary nature of the *MonNeg* assembly (and the possibility of artificially truncated gene coding sequences), an additional condition that the shortest protein across the six organisms is ≥ 85% of the length of the longest protein was also imposed, resulting in 75 families with an average of ≈27K amino acids per organism. Multiple alignments and phylogram estimation were by the ETE Toolkit `sptree_fasttree_all` /

`standard_fasttree` pipelines (56, 57). Alternatively, if *MonNeg* is ignored, there are 1,253 families before the similar length requirement, and if this requirement is loosened from 85% to 50%, 978 families with an average of ≈497K amino acids per organism proceed to the same ETE pipelines, and the resulting phylogram is very similar to that shown with just a slightly higher average rate of amino acid changes but similar proportions; this phylogram was stable when the 978 families were randomly partitioned into six groups of 163 families each. An analysis based on 16S/18S rRNA nucleotide sequences extracted from NCBI also produces a similar phylogram (but with a much lower average rate of nucleotide change). The topology of all these trees is in agreement with Leliaert*, et al.* (58).

Scatter plot showing scrambled syntenic blocks. This (and Figs. S2–S10) are similar to Fig. 2 for *CocSub* vs. *Chlore* in Blanc*, et al.* (51), but with a finer scheme for generating statistical enrichment shading as well as permutation of genome assembly sequences to emphasize enrichments. To identify statistically enriched regions, each assembly sequence is partitioned into as equal-sized pieces as possible with each piece being ≈1 Mbp (small sequences are taken whole); this induces a 2-D partitioning of the plotted area, and the number of observed gene pairs (red plus green dots) in each 2-D bin is noted. Randomized versions of the plot are then generated: for each version, the identities of all genes are shuffled in each genome and new numbers of points in each 2-D bin tallied; the *p*-value for a 2-D bin is taken as the fraction of times the random tally is larger than the observed tally over 100,000 randomizations. These *p*-values are used to shade the background of 2-D bins from white (*p*-values above 0.01) to increasingly orange on a logarithmic scale to deepest orange for *p*-values near 0.00001.

The plotted order of genome assembly sequences along each axis is determined as follows. Reordering is only performed among those sequences ("large") in an assembly ≥ 0.5 Mbp long. First, consider 2-D bins with *p*-values at or below 0.01, and form a directed graph with arcs from *x*-axis large sequences to *y*-axis large sequences with arc weights given by the total number of red plus green dots in considered 2-D bins that land in the pair of sequences, deleting arcs of weight zero. Using the `Centrality` method of `FindGraphCommunities[]` in Mathematica, partition the sequences into an ordered list of clusters. Start by considering in turn those clusters that involve both genomes: find a maximal-weight matching for the subgraph of the current cluster, place ordered pairs (*x-axis assembly original sequence number*, *y-axis assembly original sequence number*) for the matching is ascending lexicographic order, and take these as the next sequences in the reordering for both genomes; if the matching does not involve all sequences in the cluster, add the leftover sequences by ascending original assembly order. Finally, after all clusters involving both genomes are processed, in each genome add in all sequences not yet included in ascending original assembly order.

Gene prediction and functional annotation

Ab initio gene models (Dataset S19) were constructed with AUGUSTUS 3.0.3 using default parameters except where noted as follows. PASA 2.0.2 (50) was used to extract a training set of 6,576 genes from the assembled transcriptome. Prediction hints for AUGUSTUS were created by aligning the transcriptome to the genome with BLAT 35. Functional annotations were generated from protein translations of predicted gene models. For example, BLAST2GO 6.0 (59) was used to associate Gene Ontology (GO, 60) terms as well as brief textual descriptions to genes. To generate protein domain/family annotations, protein translations were scanned against PfamA release 29 with HMMER 3.1b2 (61). Additional GO associations were derived using the Pfam2GO translation table from EMBL-EBI (62). All functional enrichment analyses were based on hypergeometric statistical tests using the annotations of the entire genome as background.

Astaxanthin-deficient mutants

A non-targeted forward genetics screen generated astaxanthin-deficient mutants. Cells were grown to log phase (2–5 × 10^6 cells/mL), subjected to ultraviolet radiation (80,000 μjoules), and plated onto selection media (proteose media with 28 mM glucose). The selection media enhances the production of astaxanthin, which causes the cells to become pink; therefore green colonies were selected as astaxanthin candidate mutants. The lack of astaxanthin production was confirmed by HPLC pigment analysis. To analyze pigments, cells were scraped from plates and homogenized with acetone and lysing matrix D for 2× 60 s with the FastPrep-24 (6.5 m s^{-1}, MP Biomedical). The cell debris was pelleted by centrifugation (20,000 *g* for 3 min) and the supernatant was removed. To ensure complete extraction, another aliquot of acetone was added to the cell debris pellet and the extraction process was repeated; pigments were determined by HPLC as previously described (63). To sequence the β-carotene ketolase gene from *C. zofingiensis* wild type and astaxanthin mutants, a series of synthetic primers (Table S5) were used to amplify overlapping fragments of genomic DNA. Sequences were assembled using Lasergene MegAlign (DNASTAR) and putative point mutations were identified.

Liquid cultures of wild type and astaxanthin mutants were grown until log phase under medium light (100 μmol photons m^{-2} s^{-1}) and then high light treatment cultures were moved to 400–450 μmol photons m^{-2} s^{-1} for 10 days. Replicates (N = 3 or 4) were harvested by centrifugation and the cell pellet was frozen in liquid nitrogen. Pigment determination was conducted as described above. Pigment concentrations were tested for assumptions of normality and homoscedasticity, and data were log-transformed accordingly prior to analyses. ANOVA was used to test the effects of high light. For all significant factors in the ANOVA tests, post-hoc Tukey-Kramer HSD pairwise comparisons were used to test which groups were significantly different. α-carotene concentrations were not normally distributed and the Kruskal-Wallis non-parametric test was used instead to evaluate statistical differences. Statistical differences were reported significant at the $\alpha = 0.05$ level.

SI Figures

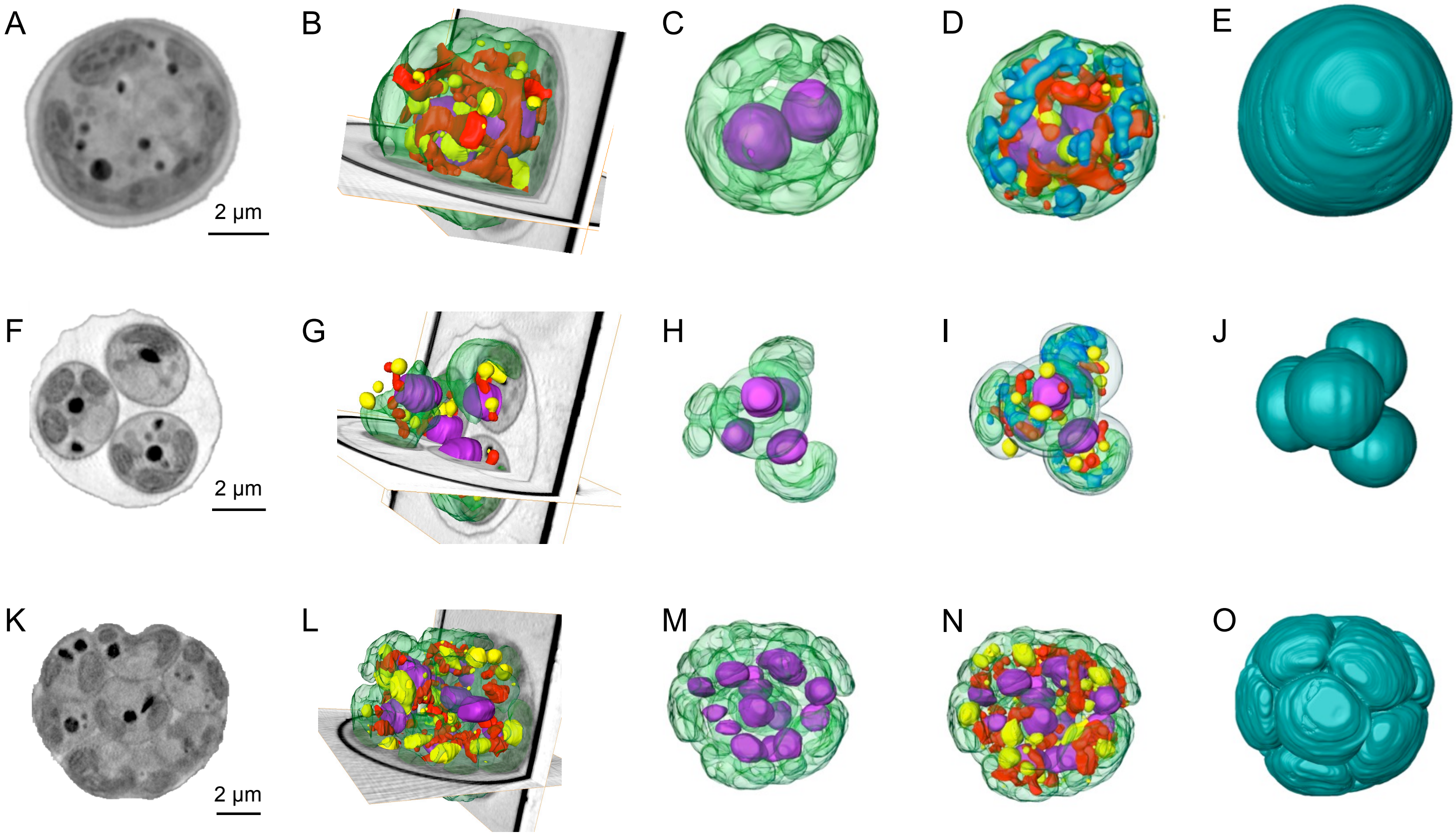


Fig. S1. ***Chromochloris zofingiensis*** **cell morphology.** Soft X-ray tomography of reconstructed cell with segmented nucleus (purple), chloroplast (green), mitochondria (red), lipids (yellow), and starch granules (blue). Cellular structures of a dividing cell with two nuclei (A–E), cell dividing into 4 cells (F–J), and dividing into 16 cells (K–O). For each cell, a representative orthoslice of the reconstructed cell (A, F, K), 3-D segmentation over two orthogonal orthoslices (B, G, L), segmented chloroplast and nucleus (C, H, M), fully segmented cell (D, I, N), and cell walls (E, J, O) are shown. Movies S1–S5 show transmission of a single cell and 3-D reconstruction of cells.

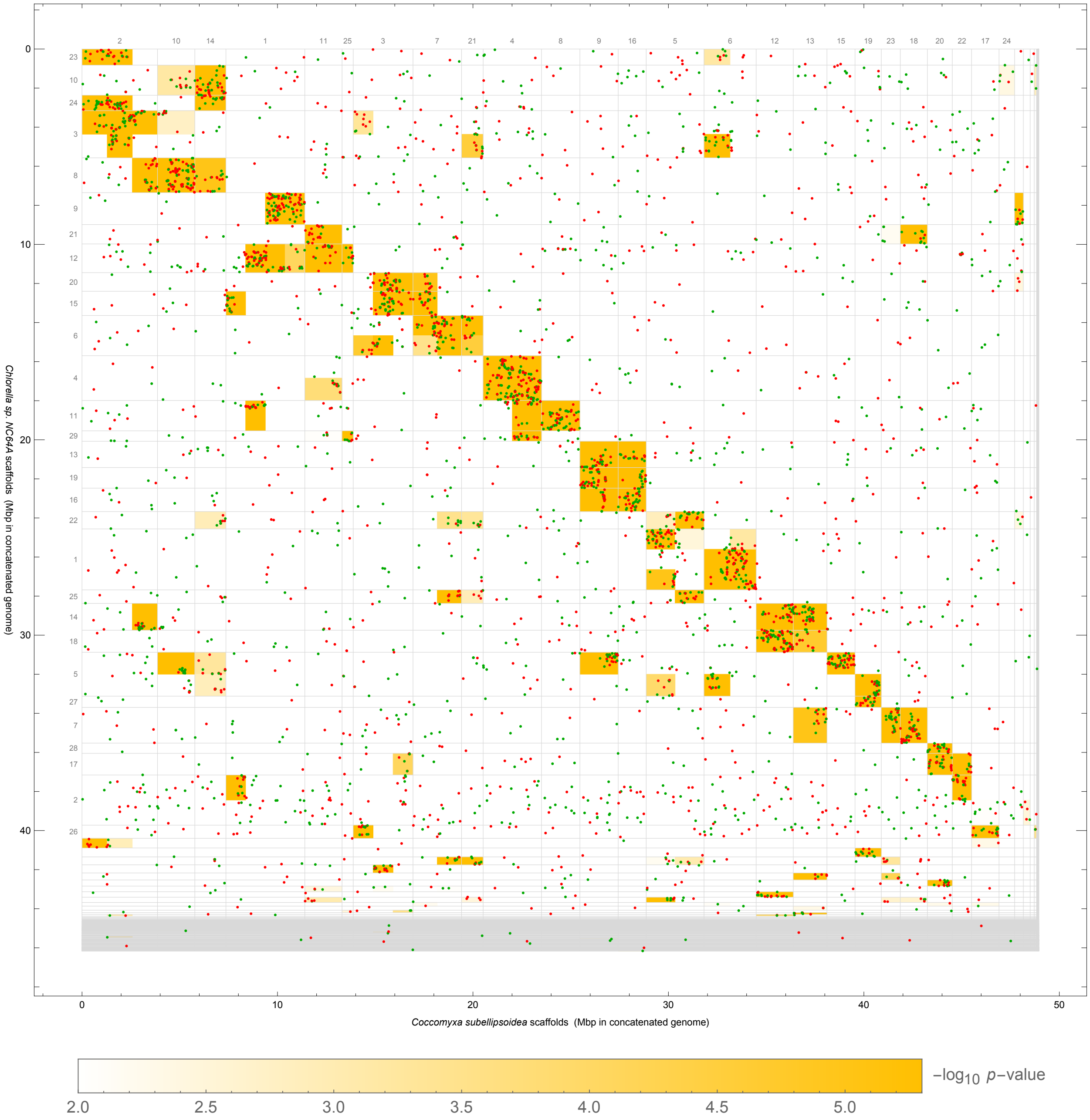


Fig. S2. Gene-level synteny for *Coccomyxa subellipsoidea C*-169 vs. *Chlorella* sp. NC64A in the style of main text Fig. 3C.

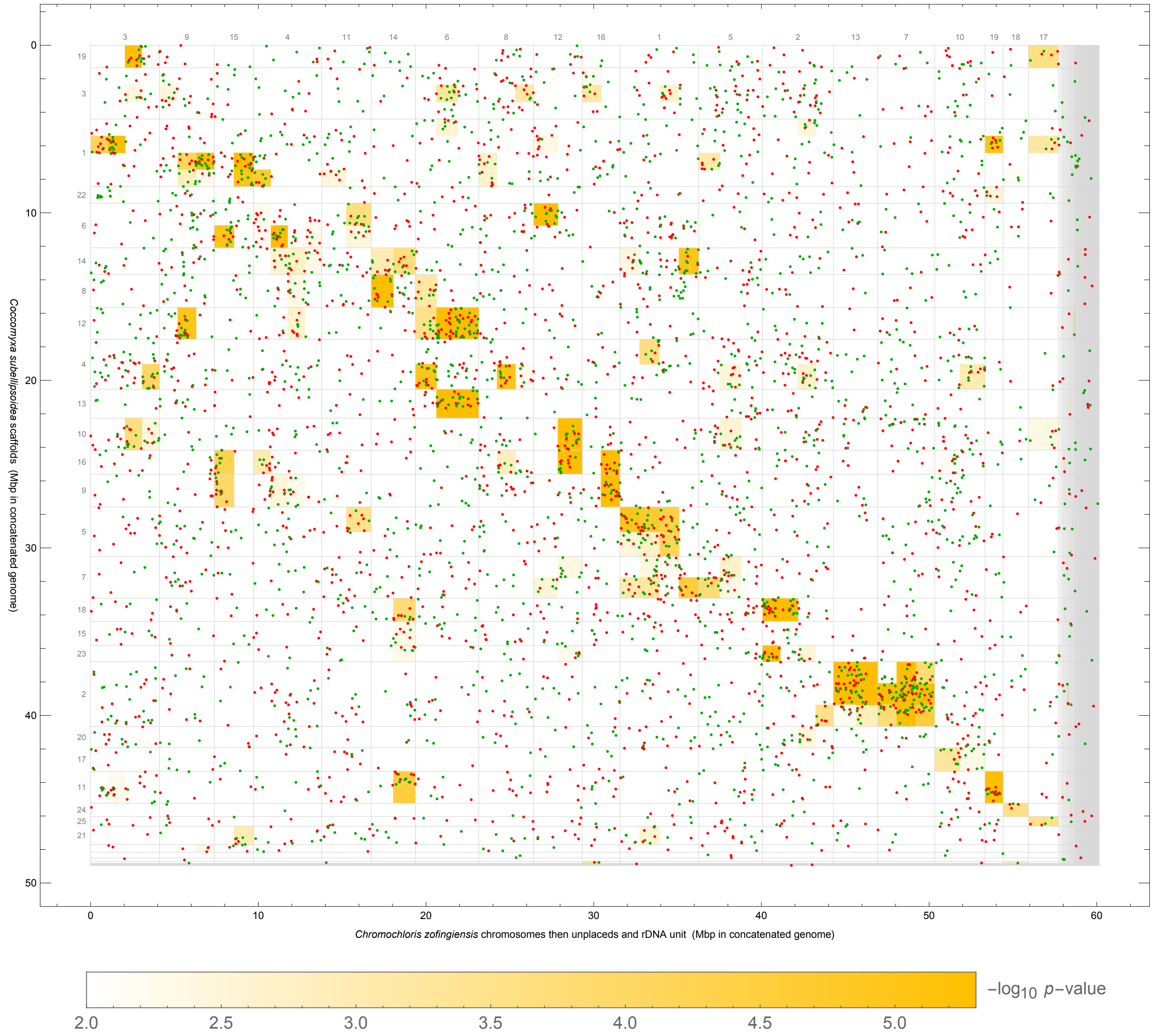


Fig. S3. Gene-level synteny for *Chromochloris zofingiensis* vs. *Coccomyxa subellipsoidea C*-169 in the style of main text Fig. 3C.

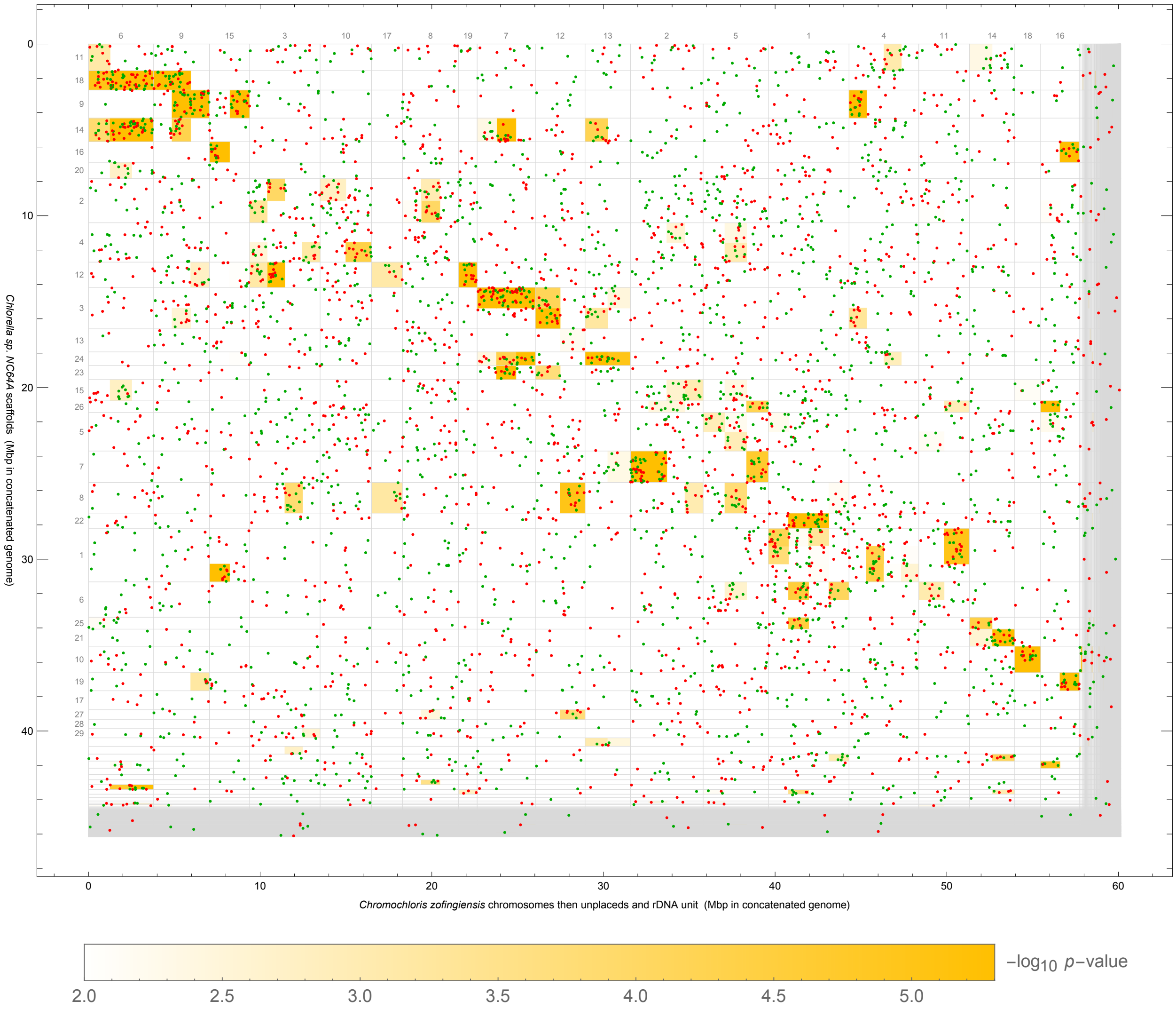


Fig. S4. Gene-level synteny for *Chromochloris zofingiensis* vs. *Chlorella* sp. NC64A in the style of main text Fig. 3C.

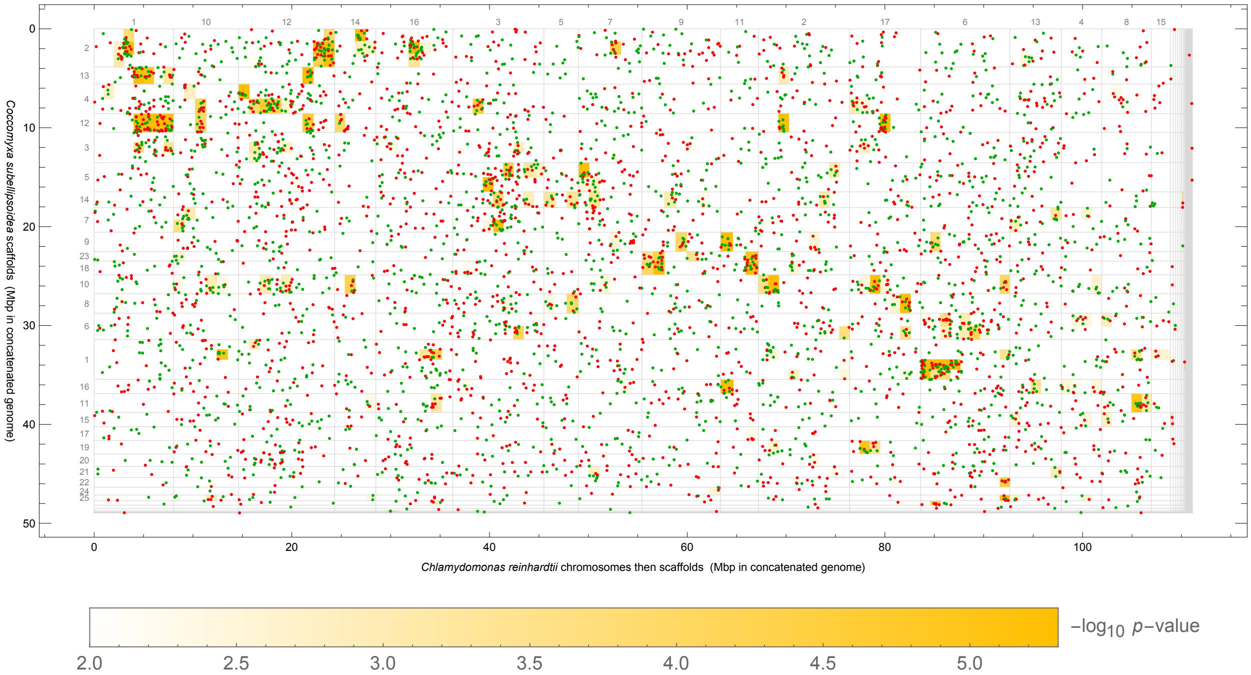


Fig. S5. Gene-level synteny for *Chlamydomonas reinhardtii* vs. *Coccomyxa subellipsoidea C*-169 in the style of main text Fig. 3C.

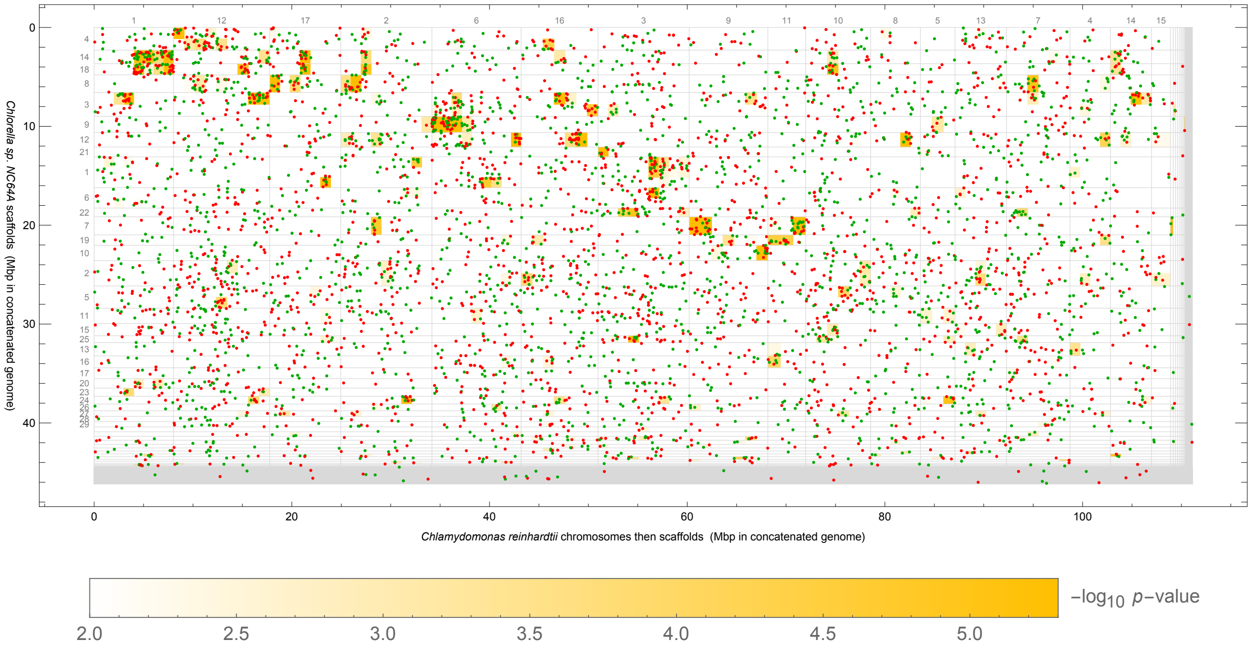


Fig. S6. Gene-level synteny for *Chlamydomonas reinhardtii* vs. *Chlorella* sp. NC64A in the style of main text Fig. 3C.

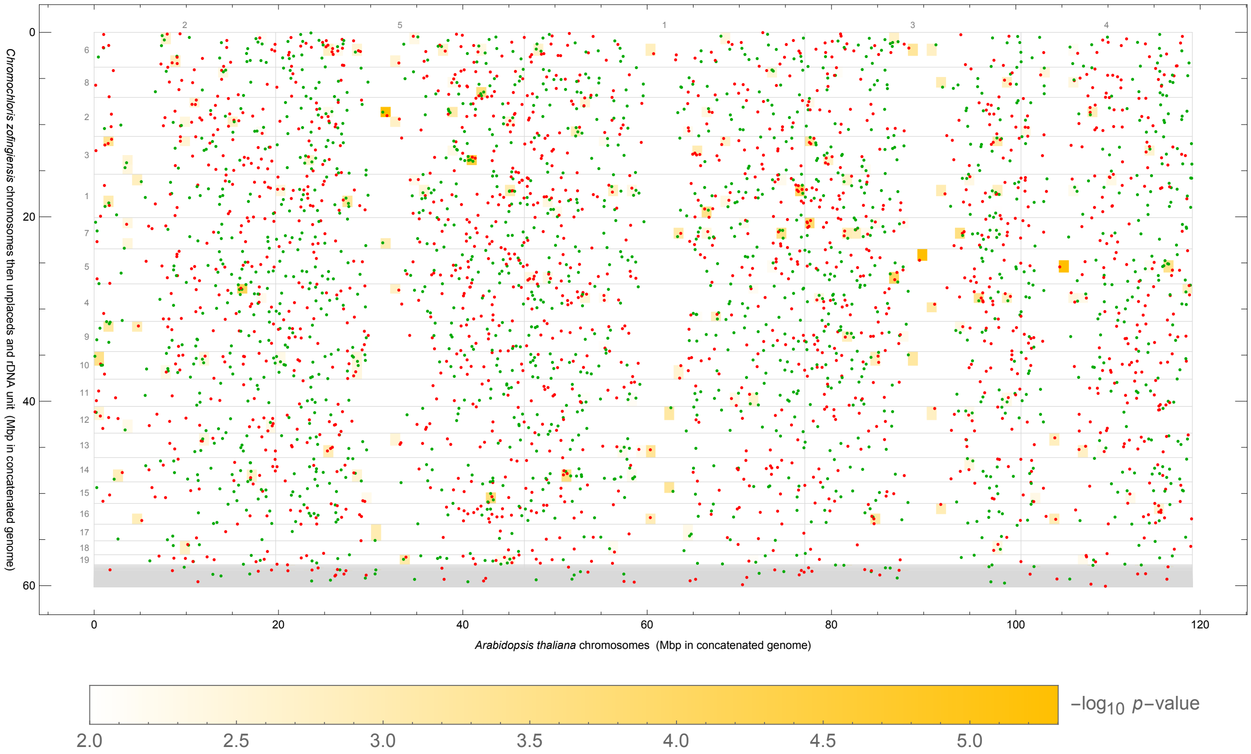


Fig. S7. Gene-level synteny for *Arabidopsis thaliana* vs. *Chromochloris zofingiensis* in the style of main text Fig. 3C.

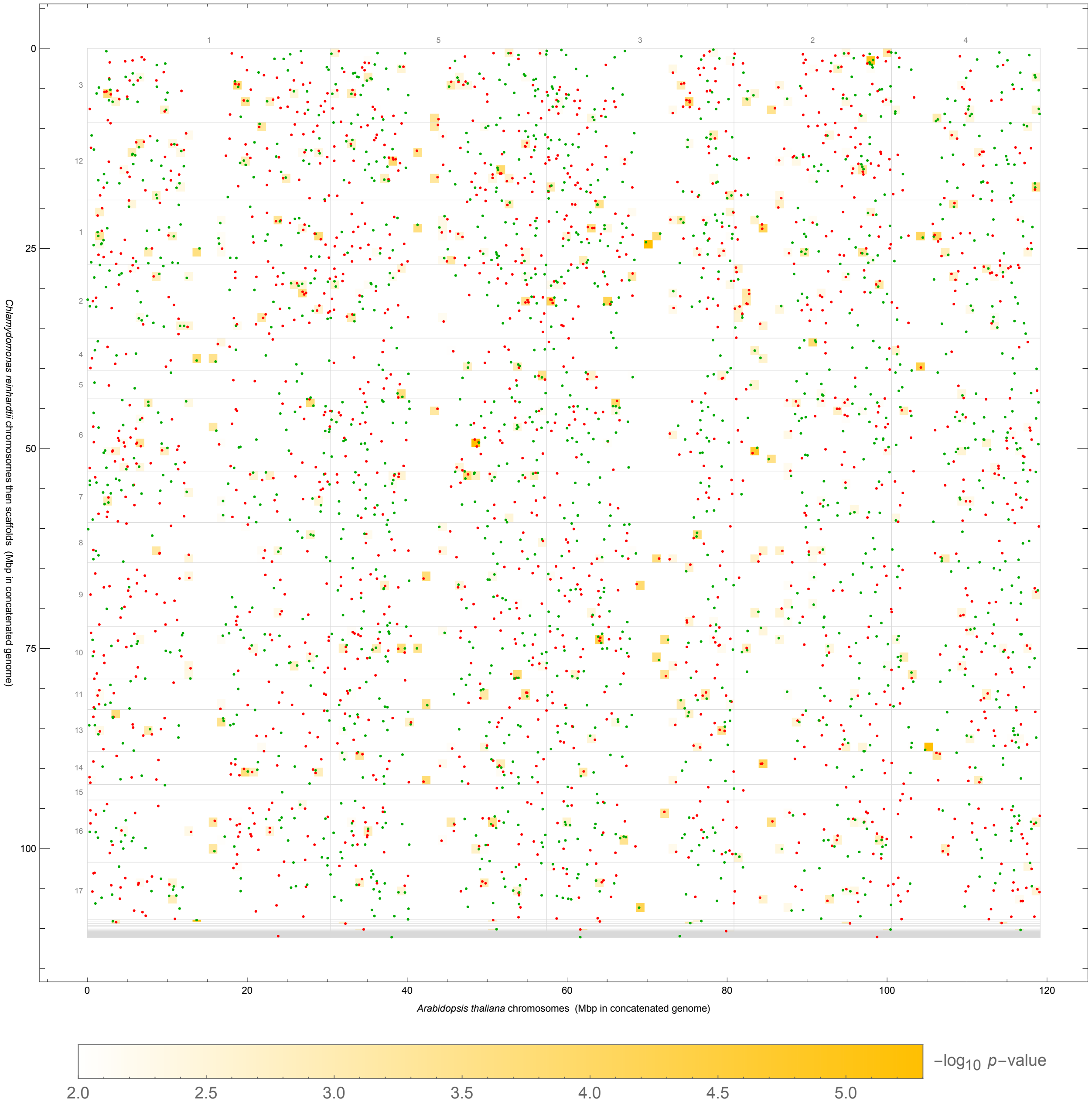


Fig. S8. Gene-level synteny for *Arabidopsis thaliana* vs. *Chlamydomonas reinhardtii* in the style of main text Fig. 3C.

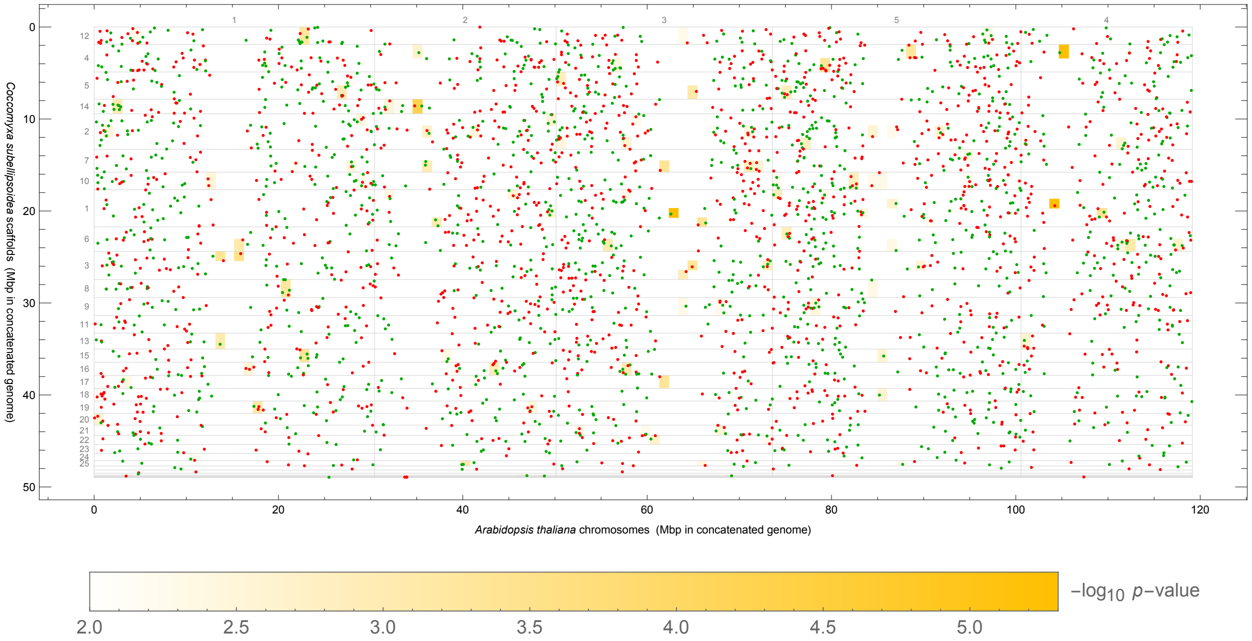


Fig. S9. Gene-level synteny for *Arabidopsis thaliana* vs. *Coccomyxa subellipsoidea C*-169 in the style of main text Fig. 3C.

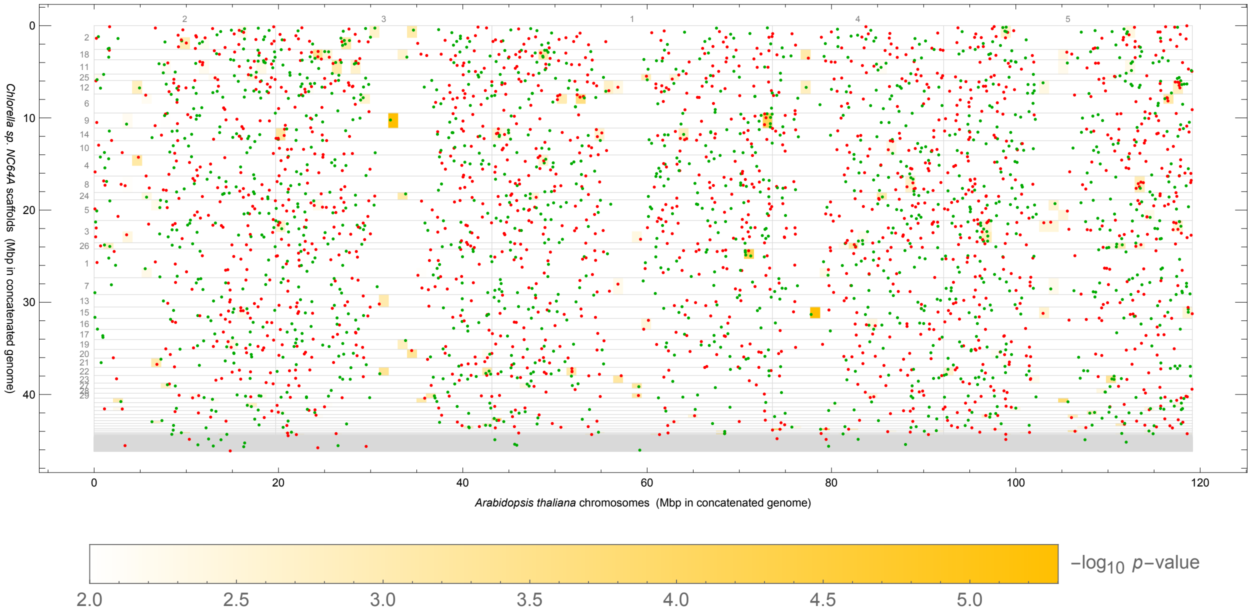


Fig. S10. Gene-level synteny for *Arabidopsis thaliana* vs. *Chlorella* sp. NC64A in the style of main text Fig. 3C.

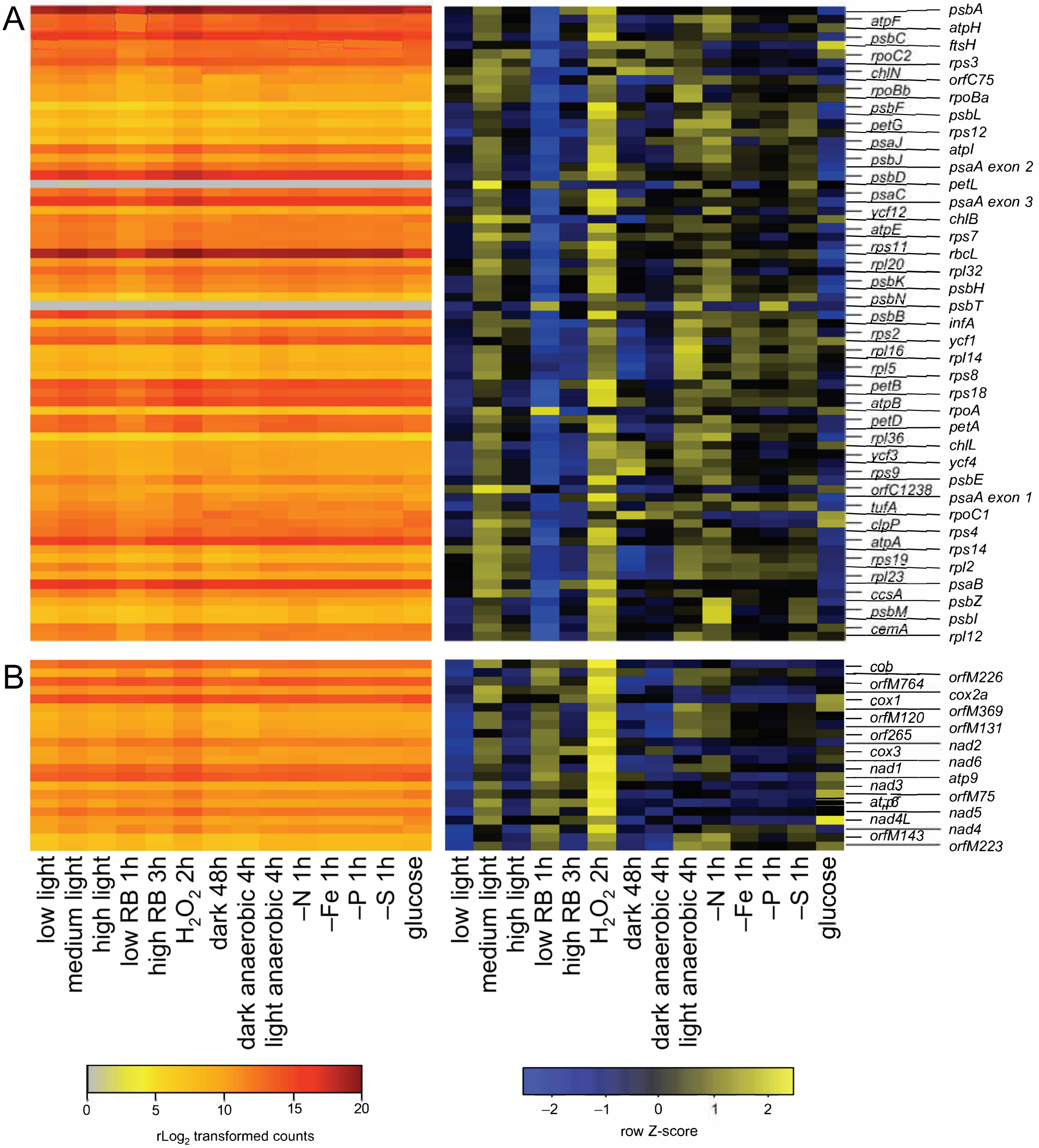

A
atpF
psbC
rpoC2
chlN
rpoBb
psbF
petG
psaJ
psbJ
psbD
psaC
ycf12
atpE
rps11
rpl20
psbK
psbN
psbB
rps2
rpl16
rpl5
petB
atpB
petD
rpl36
ycf3
rps9
orfC1238
tufA
clpP
atpA
rps19
rpl23
ccsA
psbM
cemA
psbA
atpH
ftsH
rps3
orfC75
rpoBa
psbL
rps12
atpI
psaA exon 2
petL
psaA exon 3
chlB
rps7
rbcL
rpl32
psbH
psbT
infA
ycf1
rpl14
rps8
rps18
rpoA
petA
chlL
ycf4
psbE
psaA exon 1
rpoC1
rps4
rps14
rpl2
psaB
psbZ
psbI
rpl12
B
cob
orfM764
cox1
orfM120
orf265
cox3
nad1
nad3
atp6
nad4L
orfM143
orfM226
cox2a
orfM369
orfM131
nad2
nad6
atp9
orfM75
nad5
nad4
orfM223
low light
medium light
high light
low RB 1h
high RB 3h
H_2O_2 2h
dark 48h
dark anaerobic 4h
light anaerobic 4h
–N 1h
–Fe 1h
–P 1h
–S 1h
glucose
low light
medium light
high light
low RB 1h
high RB 3h
H_2O_2 2h
dark 48h
dark anaerobic 4h
light anaerobic 4h
–N 1h
–Fe 1h
–P 1h
–S 1h
glucose
0
5
10
15
20
rLog$_2$ transformed counts
−2
−1
0
1
2
row Z-score

Fig. S11. Chloroplast and Mitochondrial RNA-Seq expression profiles across a range of growth conditions. RNA-Seq was performed on total RNA collected from *C. zofingiensis* grown under 14 different conditions including nutrient deprivation, oxidative stress, heterotrophic growth, and a range of light intensities. Total RNA depleted of rRNA was used so as to capture the non-polyadenylated transcripts expressed in the chloroplast and mitochondria. (A) Chloroplast gene expression. On the left, transcript abundances of the 73 chloroplast genes were quantified in terms of regularized $\log_2$-transformed counts and plotted as a heatmap for each of the 14 conditions. On the right, each gene was normalized across the 14 conditions and the resulting *z*-scores were plotted. (B) Mitochondrial gene expression. The 22 chloroplast genes were analyzed as in panel A.

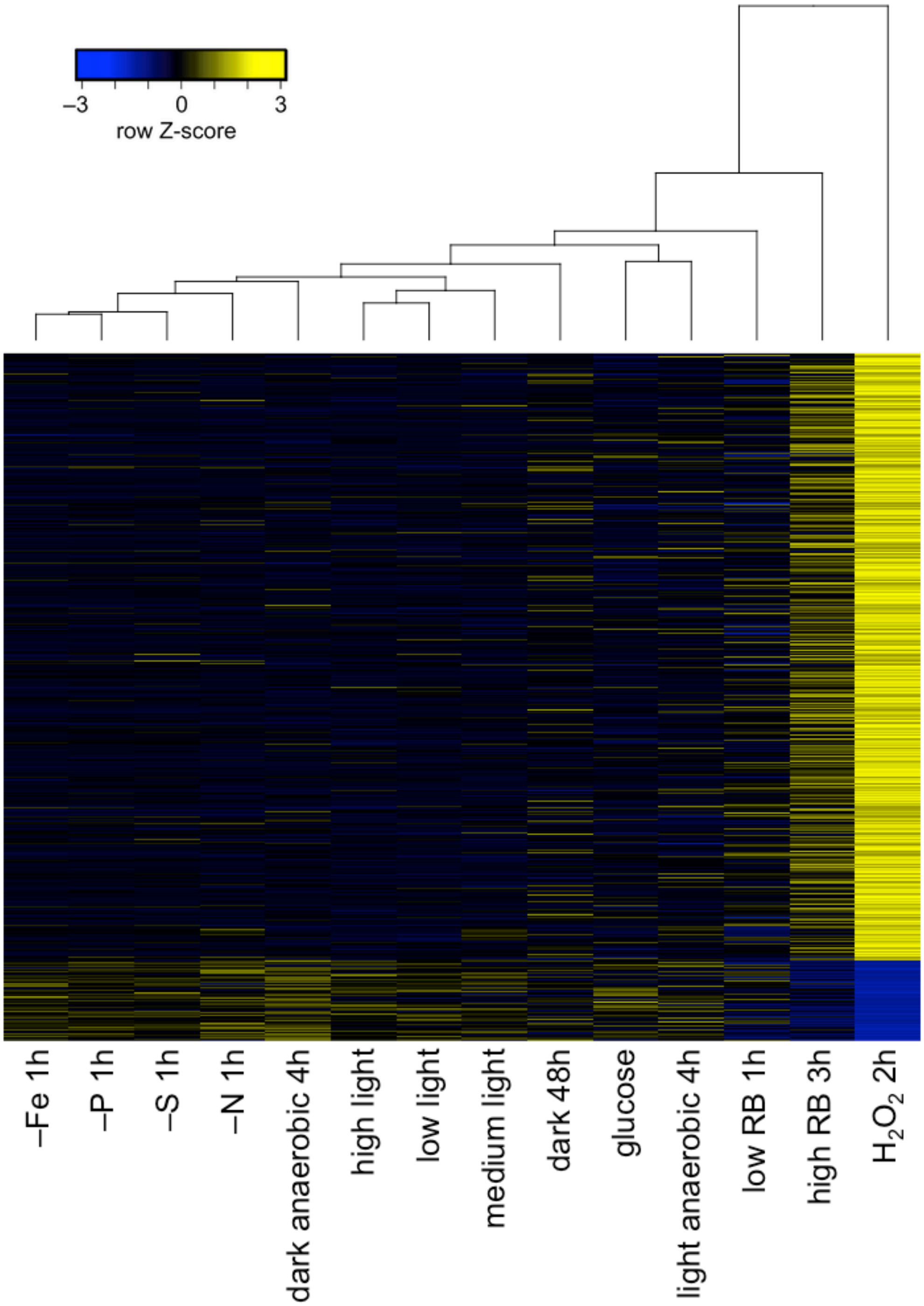


Fig. S12. Genes affected by H_2O_2 stress. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the H_2O_2-treated sample were selected and plotted as a clustered heatmap (*N* = 3934). The dendrogram above indicates the relationship between the 14 conditions.

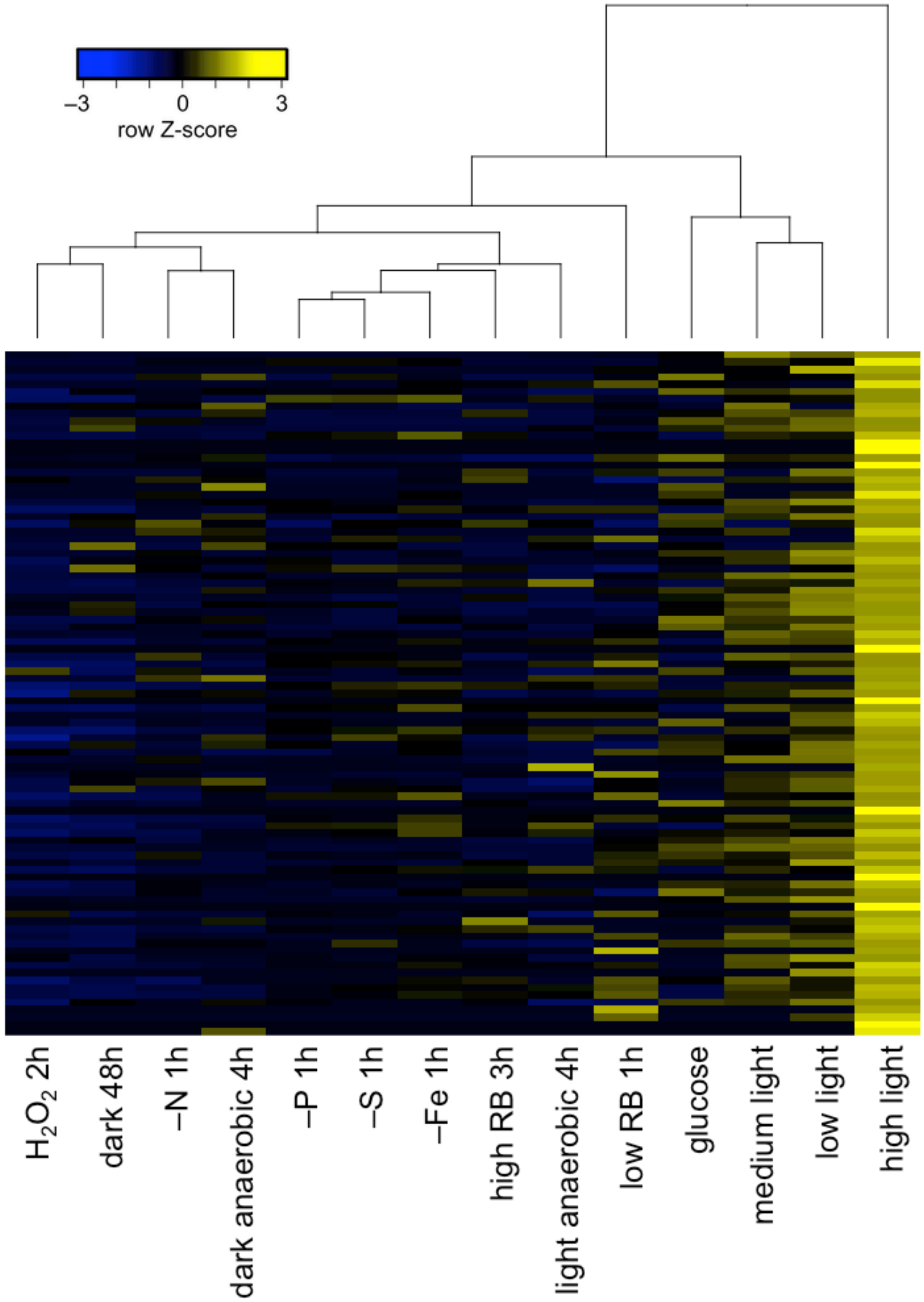


Fig. S13. Genes affected by high light. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the high light sample were selected and plotted as a clustered heatmap (*N* = 93). The dendrogram above indicates the relationship between the 14 conditions.

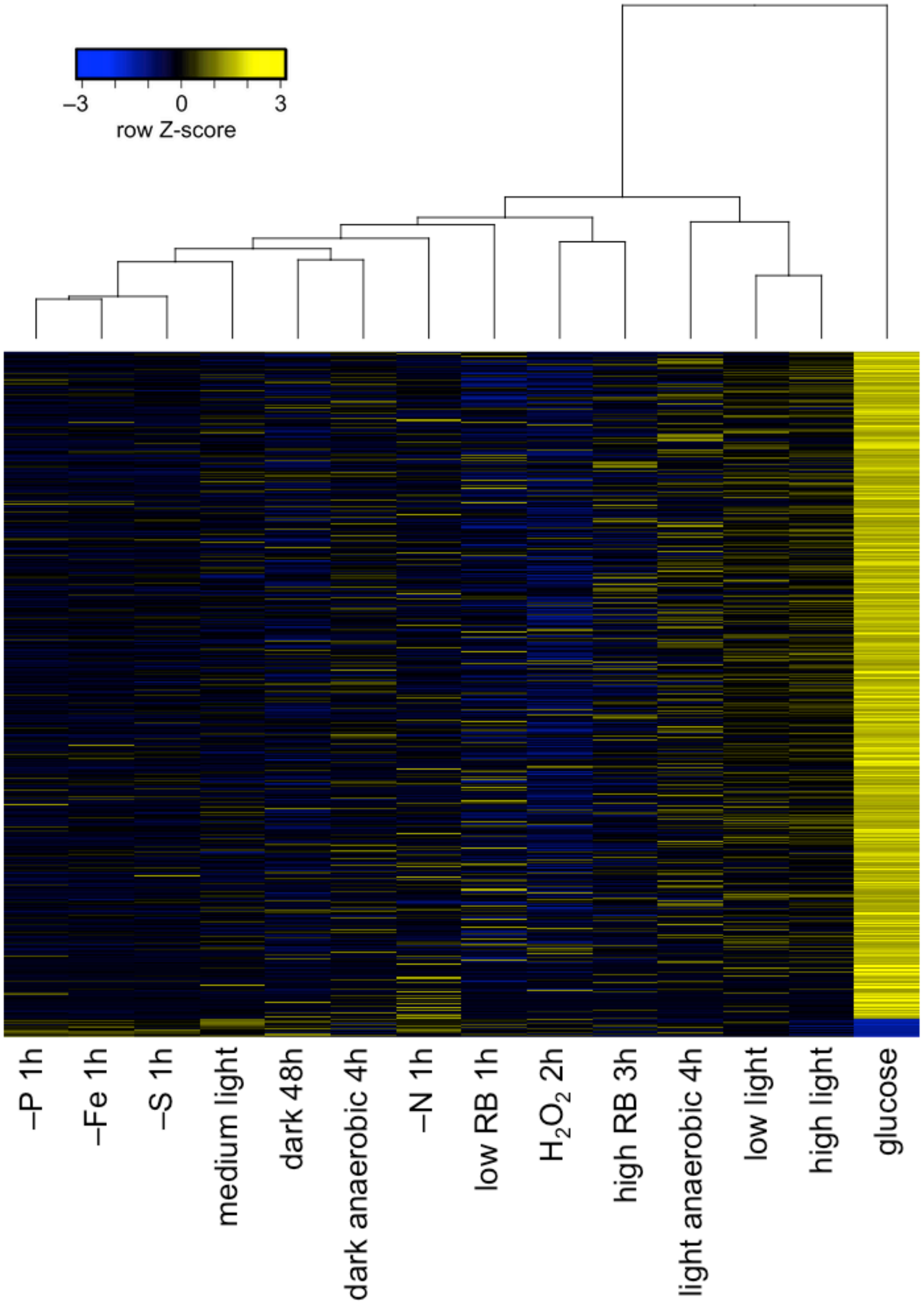


Fig. S14. Genes affected by glucose. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the glucose sample were selected and plotted as a clustered heatmap (*N* = 853). The dendrogram above indicates the relationship between the 14 conditions.

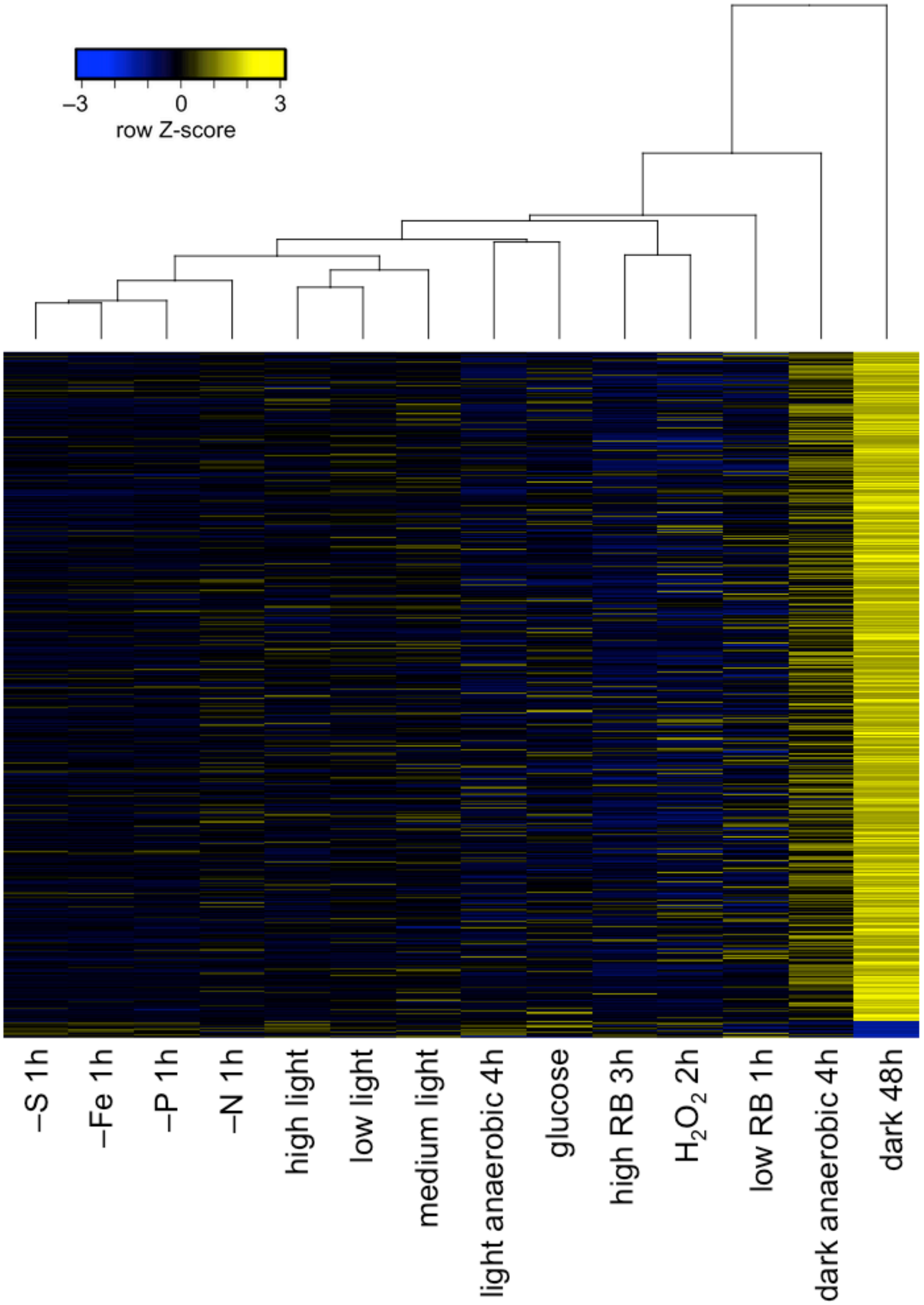


Fig. S15. Genes affected by growth in the dark. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the sample grown in the dark were selected and plotted as a clustered heatmap (*N* = 831). The dendrogram above indicates the relationship between the 14 conditions.

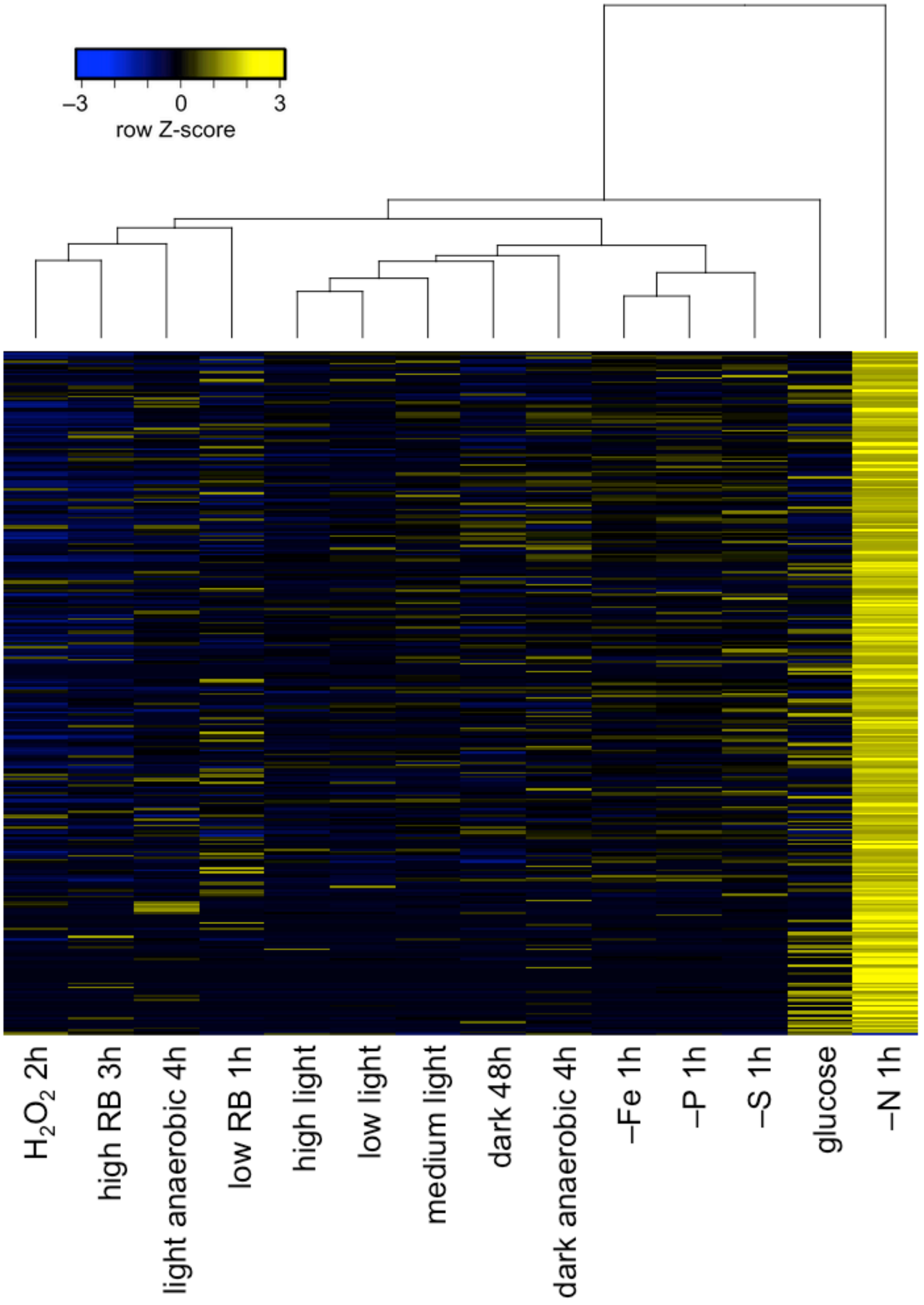


Fig. S16. Genes affected by growth in minus nitrogen (–N). Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the sample grown without N were selected and plotted as a clustered heatmap (*N* = 338). The dendrogram above indicates the relationship between the 14 conditions.

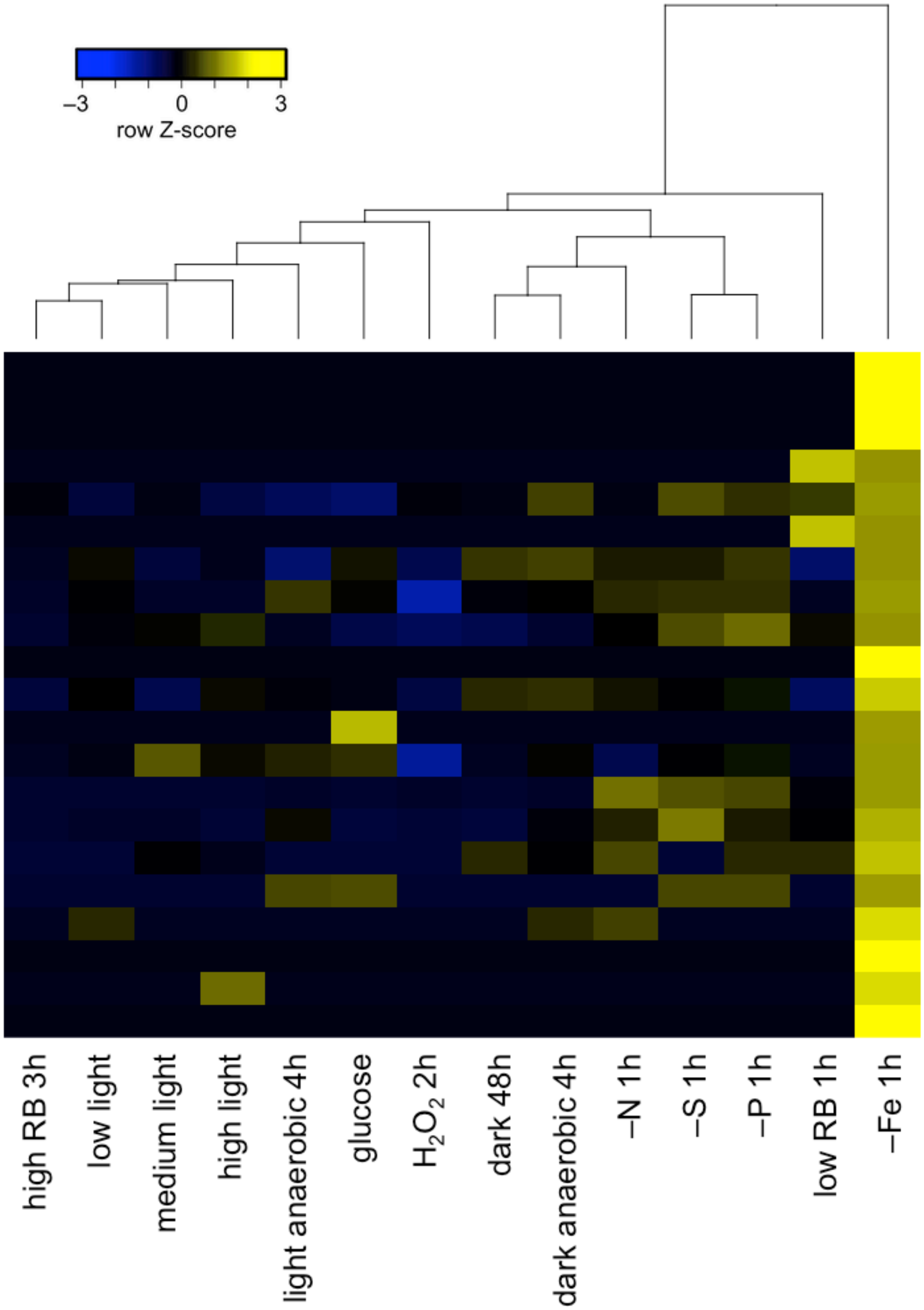


Fig. S17. Genes affected by growth in minus iron (–Fe). Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the sample grown without Fe were selected and plotted as a clustered heatmap (*N* = 21). The dendrogram above indicates the relationship between the 14 conditions.

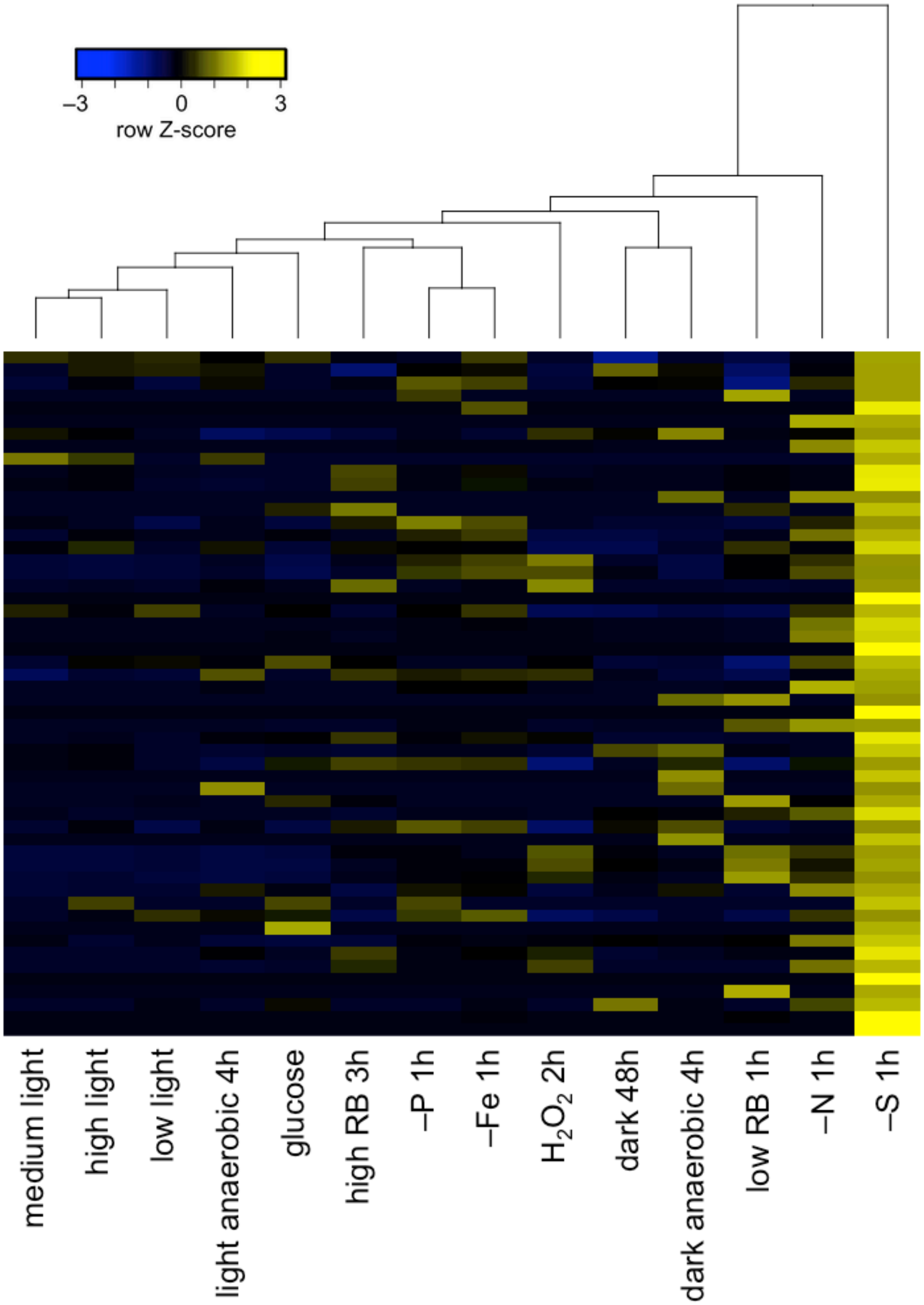


Fig. S18. Genes affected by growth in minus sulfur (–S). Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the sample grown without S were selected and plotted as a clustered heatmap (*N* = 54). The dendrogram above indicates the relationship between the 14 conditions.

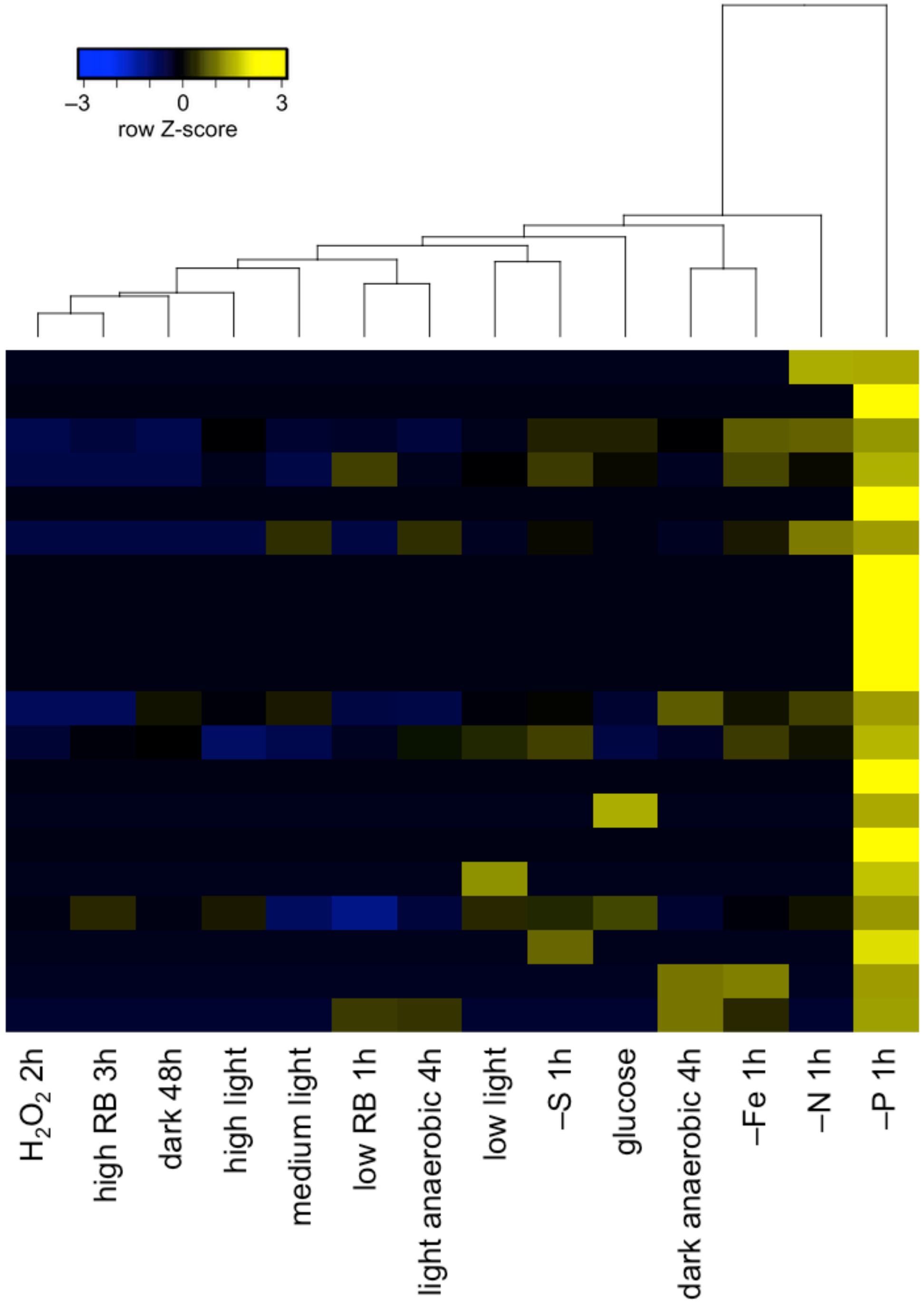


Fig. S19. Genes affected by growth in minus phosphorus (–P). Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the sample grown without P were selected and plotted as a clustered heatmap (*N* = 20). The dendrogram above indicates the relationship between the 14 conditions.

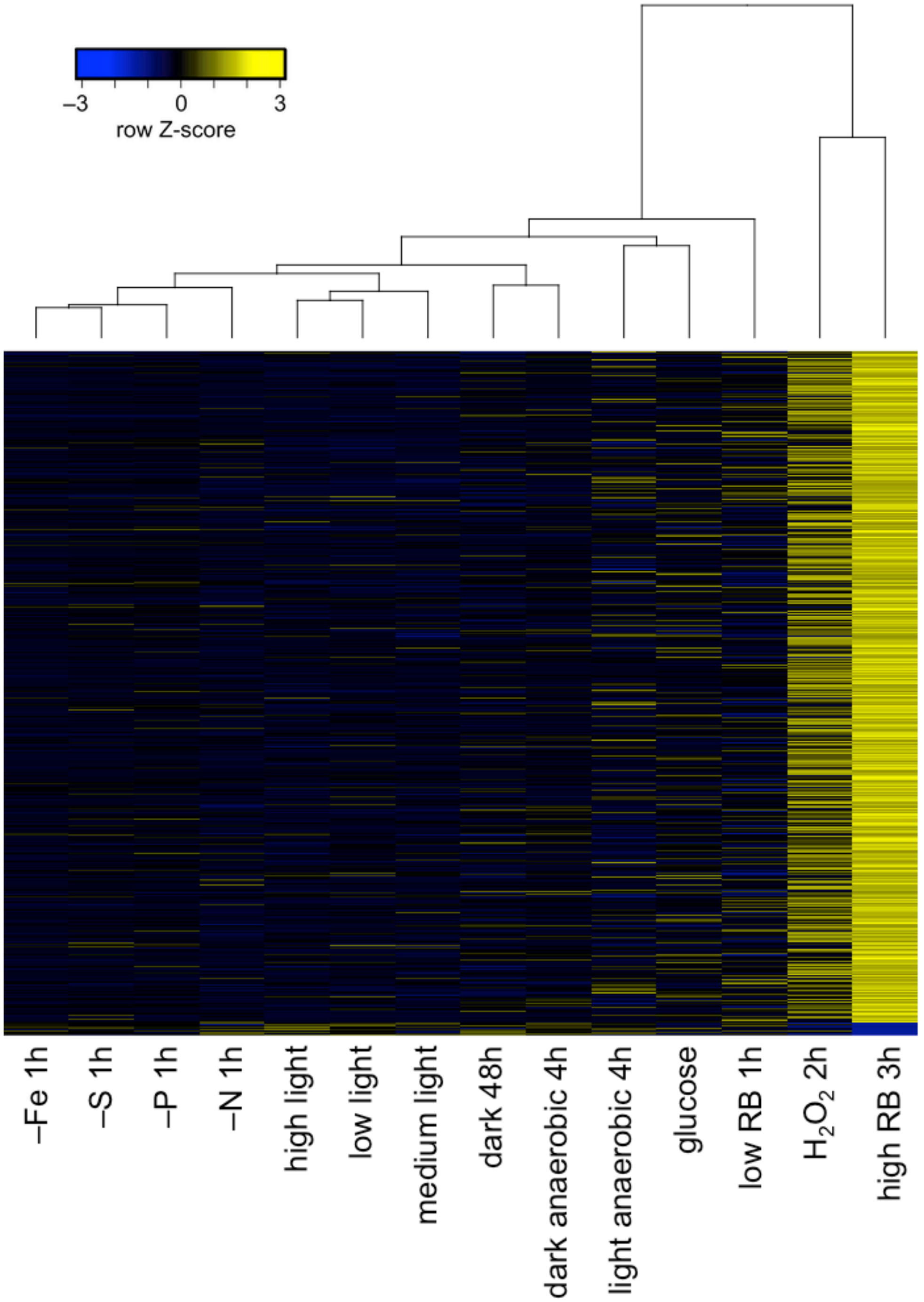


Fig. S20. Genes affected by growth in high rose bengal. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the high rose bengal treated sample were selected and plotted as a clustered heatmap (*N* = 1,477). The dendrogram above indicates the relationship between the 14 conditions.

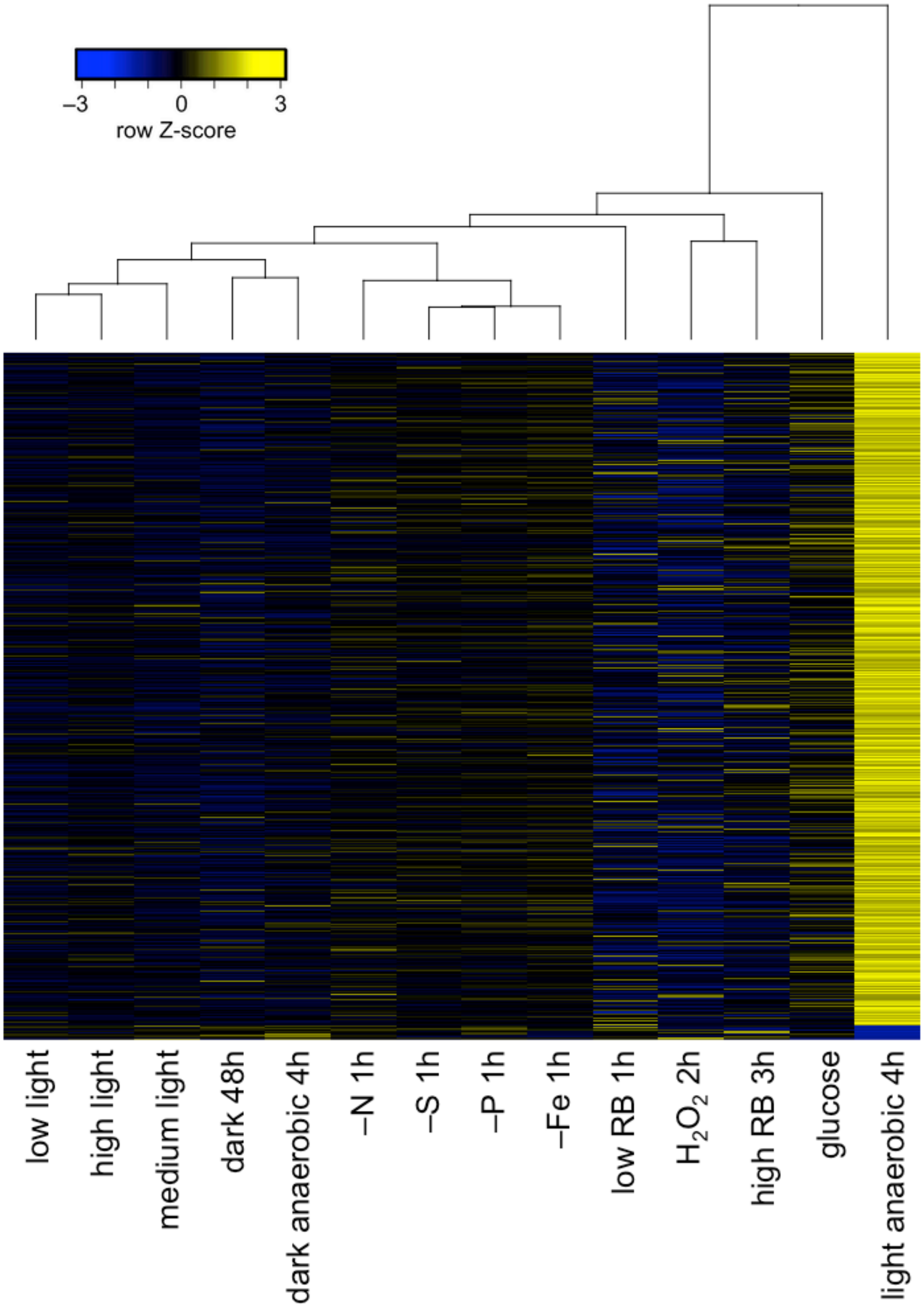


Fig. S21. Genes affected by anaerobic growth. Transcript abundance estimates for all genes were normalized across 14 different conditions. Those with a *z*-score > 2 or < –2 in the anaerobic growth sample were selected and plotted as a clustered heatmap (*N* = 1,308). The dendrogram above indicates the relationship between the 14 conditions.

SI Tables

Table S1. Features of the *Chromochloris zofingiensis* genome in comparison to selected previously sequenced genomes. The *C. zofingiensis* genome was compared to four other green algal genomes (*Chlamydomonas reinhardtii, Coccomyxa subellipsoidea* C-169, *Chlorella sp.* NC64A, and *Monoraphidium neglectum*) and the model plant *Arabidopsis thaliana*. Quantities were generally computed with uniform rules applied to most recently available genome assemblies and annotation releases (SI Appendix, SI Text).

	***Coccomyxa subellipsoidea* C-169**	***Chromochloris zofingiensis***	***Arabidopsis thaliana***	***Chlamydomonas reinhardtii***	***Chlorella sp.* NC64A**	***Monoraphidium neglectum***
Nuclear genome	JGI Phytozome 2.0 assembly and gene models	This work ("ChrZofV5"): chroms. + unplaced + 24x copies of rDNA as single contig	TAIR10 assembly and gene models	JGI Phytozome 5.5 assembly and gene models	JGI release 2014-08-18 assembly and "best genes" models only	NCBI KK100223 thru KK106940 version .1 sequences and gene models
Sequenced genome size	49 Mbp	57 Mbp	119 Mbp	107 Mbp	42 Mbp	67 Mbp
Sequenced genome presentation	29.5 Mbp in 12 contiguous chroms., 19.1 Mbp in 16 contiguous chrom. arms (pairing not known for half), 333 Kbp in 17 unplaced contigs	54.4 Mbp in 19 chroms. (4 ctg. + 15 scaf.), 2.4 Mbp in 198 unplaceds (171 ctg. + 27 scaf.), 9.7 Kbp in 1 canonical rDNA unit contig	119.0 Mbp in 5 scaffolded chromosomes	105.1 Mbp in 17 scaffolded chromosomes, 2.0 Mbp in 37 unplaceds (15 contigs + 22 scaffolds)	41.9 Mbp in 216 unplaced scaffolds + 322 Kbp in 198 unplaced contigs	43.4 Mbp in 3,257 unplaced scaffolds + 23.7 Mbp in 3,461 unplaced contigs
Sequenced genome: total # of stretches of pure "A"/"C"/"G"/"T" basepairs	**45**	**296**	**359**	**1,547**	**3,957**	**12,074**
Genome project primary initial strategy, average basepair coverage at earliest stage	Sanger WGS, ≈12x	HiSeq PE100, ≈460x	BAC/P1/TAC, complex	Plasmid/fosmid, ≈13x	Sanger WGS, ≈9x	MiSeq PE250, ≈49x
Scaffold N50 (taking genome size as sum of scaffolds as-are)	chromosomes/arms	chromosomes	chromosomes	chromosomes	1,470 Kbp	16 Kbp
Contig N50 (taking genome size as sum of contigs as-are)	1,960 Kbp	1,444 Kbp	10,898 Kbp	215 Kbp	28 Kbp	9 Kbp
Number of chromosomes	20 (asm. subtelo./PFGE/Southerns)	19 (optical map)	5 (incontrovertible)	17 (linkage groups)	12 (PFGE/asm. subtelo.)	unknown, no estimate
Percent G+C in sequenced genome	53%	51%	36%	64%	67%	65%
Basepairs called as coding (in any transcript model) in sequenced genome	25%	39%	28%	37%	32%	26%
Percent G+C in basepairs called as coding (in any transcript model)	61%	53%	44%	70%	69%	70%
Number of called protein-coding gene loci (collapsing transcript forms)	9,629	15,274	27,206	17,741	9,791	16,734
Number of "complete" called protein-coding gene loci (collapsing transcripts)	8,815	15,194	27,197	17,685	8,509	14,268
Number of rDNA units estimated to exist in true monoploid genome	unknown, no estimate	≈24 on chrom. 13	≈375 NOR2 + 375 NOR4	≈840 total chr. 1+7+15	unknown, no estimate	≈23 total
Number of tRNAs called in sequenced genome	91	75	631	259	43	38
Taking a single representative transcript model per called protein-coding gene locus:						
Number of amino acids: average	427 aa	482 aa	407 aa	736 aa	456 aa	348 aa
Number of amino acids: median	333 aa	347 aa	350 aa	500 aa	358 aa	265 aa
Number of exons containing coding sequence: average	8.1	5.0	5.2	8.5	8.3	5.0
Number of exons containing coding sequence: median	7	4	3	7	7	4
Exon length (restricted to coding sequence): average	159 nt	291 nt	237 nt	261 nt	166 nt	207 nt
Exon length (restricted to coding sequence): median	144 nt	194 nt	133 nt	133 nt	119 nt	129 nt
Intron length (between exons with coding sequence): average	284 nt	267 nt	157 nt	269 nt	207 nt	302 nt
Intron length (between exons with coding sequence): median	246 nt	260 nt	98 nt	228 nt	171 nt	254 nt
Percentage with at least one intron (between exons with coding sequence)	94%	82%	76%	92%	98%	82%
% of seq. basepairs RepeatMasker'd with Repbase Update "eukaryotic"	2.2%	3.7%	18.0%	17.8%	8.9%	8.1%
% of seq. basepairs RepeatMasker'd with RepeatModeler	5.7%	4.5%	16.9%	21.8%	12.3%	8.9%
% of seq. basepairs RepeatMasker'd with RepeatModeler + Repbase Update "eukaryo	6.0%	5.9%	20.8%	23.0%	12.6%	9.3%
Chloroplast genome	NCBI NC_015084.1 (with one gap and no large inv. rpt.) and annots.	This work ("ChrZofV5")	NCBI AP000423.1 sequence and annotations	NCBI FJ423446.1 sequence and annotations	NCBI KP271969.1 sequence (no large inv. rpt.) and annotations	NCBI CM002678.1 seq., paper's annotations
Sequenced genome size	176 Kbp	181 Kbp	154 Kbp	204 Kbp	125 Kbp	135 Kbp
Number of annotated protein-coding genes, including hypotheticals	80	71	86	67 + 1 ncRNA (tscA)	79	67
Number of annotated rRNAs	3	6	8	10	3	6
Number of annotated tRNAs	32	31	37	29	31	29
Percent G+C in sequenced genome	51%	31%	36%	34%	34%	32%
Mitochondrial genome	NCBI NC_015316.1 sequence and annotations	This work ("ChrZofV5")	NCBI JF729201.1 sequence and annotations	NCBI NC_001638.1 sequence and annotations	NCBI NC_025413.1 sequence and annotations	NCBI CM002677.1 seq. (with two gaps) and annotations
Sequenced genome size	65 Kbp	42 Kbp	367 Kbp	16 Kbp	78 Kbp	93 Kbp
Number of annotated protein-coding genes, including hypotheticals	31	22	32	8	32	17
Number of annotated rRNAs	3	6	3	14	3	0
Number of annotated tRNAs	26	24	21	3	27	23
Percent G+C in sequenced genome	53%	36%	45%	45%	28%	46%

Table S2. Homology Determination by BLAST of NCBI Gene Accessions in *Chromochloris zofingiensis.*

NCBI Accession	NCBI Gene Symbol	Record Type	Size, NT	*C.zofingiensis* gene ID	*C.zofingiensis* gene symbol	*C.zofingiensis* locus	BLAST Identities	BLAST Gaps
GQ996718.1	*BC*	gDNA	4681	Cz13g10110	none	chr13:1039758-1044449	4659/4701 (99%)	29/4701 (1%)
AY772714.1	*BKT*	gDNA	2966	Cz13g13100	*BKT1*	chr13:1335024-1337989	2960/2967 (100%)	2/2967 (0%)
FN563999.1	*LCYB*	gDNA	3683	Cz12g10170	*LCYB*	chr12:1059448-1063130	3682/3683 (99%)	0/3683 (0%)
HE664108.1	*LCYE*	gDNA	3973	Cz09g18310	*LCYE*	chr09:1898540-1902502	3963/3963 (100%)	0/3963 (0%)
KC316012.1	*NIT1*	gDNA	3448	Cz14g14210	none	chr14:1478407-1481856	3448/3450 (99%)	2/3450 (0%)
EF621406.1	*PDS*	gDNA	6713	Cz02g32280	*PDS*	chr02:3304203-3297493	6701/6717 (99%)	10/6717 (0%)
FR670784.1	*PSY*	gDNA	3561	Cz05g32220	*PSY*	chr05:3275040-3278591	3552/3553 (99%)	1/3553 (0%)
KC316010.1	*RBCS*	gDNA	2580	Cz17g13100	*RBCS*	chr17:1344226-1346748	2511/2523 (99%)	12/2523 (0%)
GQ996720.1	*SAD*	gDNA	3479	Cz04g09090	*SACPD1*	chr04:926651-930102	3424/3466 (99%)	28/3466 (1%)
HE863826.1	*ZEP*	gDNA	5687	Cz07g30060	*ZEP1*	chr07:3027178-3032856	5679/5679 (100%)	0/5679 (0%)
EU016205.1	*CHYB*	mRNA	1545	Cz12g16080	*CHYB*			
			1-534			chr12:1634792-1635325	534/534 (100%)	0/534 (0%)
			534-623			chr12:1635520-1635609	89/90 (99%)	0/90 (0%)
			614-779			chr12:1635818-1635983	166/166 (100%)	0/166 (0%)
			779-891			chr12:1636203-1636315	113/113 (100%)	0/113 (0%)
			889-1531			chr12:1636599-1637241	640/643 (99%)	0/643 (0%)
GM040753.1	"Sequence 1"	mRNA	1153	Cz13g13100	*BKT1*			
			7-210			likely cloning vector pAC106 or similar		
			209-361			chr13:1337779-1337627	153/153 (100%)	0/153 (0%)
			358-577			chr13:1337388-1337169	220/220 (100%)	0/220 (0%)
			574-749			chr13:1336856-1336681	176/176 (100%)	0/176 (0%)
			744-1150			chr13:1336336-1335930	406/407 (99%)	0/407 (0%)

Table S3: Putative *Chromochloris zofingiensis* gene models for photosynthesis-related and metabolic genes. For nuclear encoded genes, Predalgo predictions were used to determine localization (Loc.) noted using the following abbreviations: C, chloroplast; M, mitochondria; SP, secretory pathway; and O, other which could include the nucleus or cytosol. Chloroplast encoded genes are noted with CE.

Description	Gene Name	Gene ID	Loc.
Carotenoid Biosynthesis			
Geranylgeranyl Pyrophosphate Synthase	*GGPS*	Cz02g19200	C
Phytoene Synthase	*PSY*	Cz05g32220	M
Phytoene Desaturase	*PDS*	Cz02g32280	C
Zeta-Carotene Desaturase	*ZDS*	Cz10g17010	C
Zeta-Carotene Isomerase	*ZISO*	Cz10g17130	C
Carotene Isomerase	*CRTISO1*	Cz12g03260	O
Lycopene Epsilon-Cyclase	*LCYE*	Cz09g18310	C
Lycopene Beta-Cyclase	*LCYB*	Cz12g10170	O
Cytochrome P450-Type Carotene Hydroxylase (*LUT5*)	*CYP97A1*	Cz13g16110	C
	CYP97A2	Cz09g14130	O
Beta-Carotene Hydroxlase	*CHYB*	Cz12g16080	O
Cytochrome P450-Type Carotene Hydroxylase (*LUT1*)	*CYP97C*	Cz09g07100	C
Chlorophycean Violaxanthin De-Epoxidase	*CVDE*	Cz13g15030	O
Zeaxanthin Epoxidase	*ZEP1*	Cz07g30060	SP
	ZEP2	Cz05g31040	O
Neoxanthin Synthase (*ABA4*)	*NSY*	Cz15g04070	C
Beta-Ketolase	*BKT1*	Cz13g13100	O
	BKT2	Cz04g11250	O
CruP-Type Lycopene Cyclase Paralog	*CRUP*	Cz07g07080	M
NPQ			
Light-Harvesting Complex Stress Response (LI818)	*LHCSR*	Cz16g02040	C
Photosystem II Subunit S	*PSBS*	Cz03g01250	C
Chlorophyll Biosynthesis			
Glutamyl-Glutaminyl Non-Discriminatory tRNA Synthetase	*GTS1*	Cz19g08190	M
	GTS2	Cz16g14040	O
Glutamyl tRNA Reductase	*GTR*	Cz03g02090	O
Glutamate-Semialdehyde Aminotransferase	*GSA*	Cz11g23070	C
Delta-Aminolevulinic Acid Dehydratase	*ALAD*	Cz09g03180	C
Porphobilinogen Deaminase	*PBGD1*	Cz03g22200	C
	PBGD2	Cz04g12200	O
Uroporphyrinogen III Synthase	*UROS1*	Cz14g24240	O
	UROS2	Cz12g20200	SP
Uroporphyrinogen III Decarboxylase	*UROD1*	Cz14g23190	C

	UROD2	Cz16g13070	C
Coproporphyrinogen-III Oxidase	*CPOX1*	Cz06g22110	C
	CPOX2	Cz18g10090	O
	CPOX3	Cz01g34010	SP
Protoporphyrinogen IX Oxidase	*PPOX*	Cz02g42010	C
Mg-Chelatase Subunit I	*CHLI1*	Cz06g16260	O
	CHLI2	UNPLg00024	O
Mg-Chelatase Subunit H	*CHLH1*	Cz14g15070	C
	CHLH2	Cz06g36090	C
Mg-Chelatase Subunit D	*CHLD*	Cz08g03270	O
Tetrapyrrole Binding Protein	*GUN4*	Cz11g24180	O
Mg-Protoporphyrin IX Methyltransferase	*CHLM*	Cz02g24270	O
Mg-Protoporphyrin Monomethyl Ester Cyclase	*CRD1*	Cz16g08310	C
Ycf 54 Conserved Hypothetical Protein	*YCF54*	Cz19g06290	C
Divinyl Chlorophyllide *a* 8-Vinyl Reductase	*DVR*	Cz03g06230	C
Light-Dependent Protochlorophyllide Oxidoreductase	*POR*	Cz06g28160	C
Light-Independent Protochlorophyllide Oxidoreductase Complex	*chlN*	CzCPg00090	CE
	chlB	CzCPg00300	CE
	chlL	CzCPg01150	CE
Chlorophyll Synthetase	*CHLG*	Cz08g28070	M
Chlorophyllide *a* Oxygenase	*CAO1*	Cz10g28270	C
	CAO2	Cz14g04140	
Chlorophyll Degradation			
Pheophorbide *a* Oxygenase	*PAO1*	Cz08g09050	O
	PAO2	Cz06g27150	C
	PAO3	Cz14g19030	C
	PAO4	Cz14g25130	O
	PAO5	Cz16g19110	SP
Chlorophyllide *b* Reductase	*NYC1*	Cz08g00200	M
	NOL	Cz09g17200	C
Pheophytinase	*PPH1*	Cz02g13050	C
Pale Cress Etiolation Regulator	*PAC*	Cz12g26270	O
Non-Yellowing	*NYE1*	Cz10g29110	O
Heme Biosynthesis and Metabolism			
Ferrochelatse	*FC*	Cz16g21260	C
Uroporphyrin III C-Methyltransferase	*SUMT1*	Cz14g07090	O
	SUMT2	Cz05g21010	O
Heme Oxygenase	*HO*	Cz12g15230	C

	GUN2A	UNPLg00465	C
	GUN2B	UNPLg00457	C
Cytochrome Complexes and Soluble Electron Carriers			
Cyt *f*	*petA*	CzCPg01130	CE
Cyt b_6	*petB*	CzCPg01070	CE
2Fe-2S Rieske Protein	*PETC*	Cz13g12160	C
Cyt b_6f Subunit IV	*petD*	CzCPg01120	CE
Cyt b_6f Subunit G	*petG*	CzCPg00150	CE
Cyt b_6f Subunit L	*petL*	CzCPg00240	CE
Cyt b_6f Subunit VII	*PETM*	Cz07g02140	C
Cyt b_6f Subunit VIII	*PETN*	Cz16g12040	C
Cyt b_6f Subunit V	*PETO*	Cz06g16040	C
Plastocyanin	*PETE*	Cz01g35130	C
Ferredoxin	*PETF*	Cz07g28150	C
	FDX2	Cz13g08120	C
	FDX3	Cz11g03180	O
	FDX4	Cz13g08015	C
	FDX5	Cz07g16090	SP
	FDX6	Cz11g16050	O
	FDX7	Cz14g22270	C
Ferredoxin NADP Reductase	*FNR*	Cz02g17180	C
Cyt c_6	*PETJ*	Cz01g35120	O
Cyt c	*CYC4*	Cz17g14110	O
Fe-S Cluster Assembly Factor	*ISC1*	Cz08g08230	C
Cyt *c*-Type Biogenesis Factor	*CCDA1*	Cz08g31010	C
	CCB1	Cz02g33170	C
	CCB2	Cz02g07090	O
	CCB4	Cz17g14140	O
Ferredoxin-Thioredoxin Reductase, alpha Subunit	*FTRV*	Cz12g03230	C
Ferredoxin-Thioredoxin Reductase, beta Subunit	*FTRC*	Cz01g33170	O
Photosystem I			
Photosystem I Subunit A	*psaA exon1*	CzCPg01210	CE
	psaA exon2	CzCPg00200	CE
	psaA exon3	CzCPg00270	CE
Photosystem I Subunit B	*psaB*	CzCPg01360	CE
Photosystem I Subunit C	*psaC*	CzCPg00250	CE
Photosystem I Subunit D	*PSAD*	Cz05g36030	C
Photosystem I Subunit E	*PSAE*	Cz04g01180	C
Photosystem I Subunit F	*PSAF*	Cz13g06010	C

Photosystem I Subunit G	*PSAG*	Cz06g31280	C
Photosystem I Subunit H	*PSAH*	Cz12g12230	C
Photosystem I Subunit I	*PSAI*	Cz01g05275	C
Photosystem I Subunit J	*psaJ*	CzCPg00170	CE
Photosystem I Subunit K	*PSAK*	Cz06g26010	C
Photosystem I Subunit L	*PSAL*	Cz02g18180	C
Photosystem I Subunit N	*PSAN*	Cz17g10010	SP
Photosystem I Subunit O	*PSAO*	Cz15g01010	C
Photosystem II			
D1	*psbA*	CzCPg00010	CE
CP47	*psbB*	CzCPg00420	CE
CP43	*psbC*	CzCPg00040	CE
D2	*psbD*	CzCPg00210	CE
Cytochrome b_{559} Subunit E	*psbE*	CzCPg01190	CE
Cytochrome b_{559} Subunit F	*psbF*	CzCPg00130	CE
Photosystem II Subunit H	*psbH*	CzCPg00390	CE
Photosystem II Subunit I	*psbI*	CzCPg01420	CE
Photosystem II Subunit J	*psbJ*	CzCPg00190	CE
Photosystem II Subunit K	*psbK*	CzCPg00380	CE
Photosystem II Subunit L	*psbL*	CzCPg00140	CE
Photosystem II Subunit M	*psbM*	CzCPg01390	CE
Photosystem II Subunit N	*psbN*	CzCPg00400	CE
Photosystem II Subunit O	*PSBO*	Cz16g19040	C
Photosystem II Subunit P	*PSBP*	Cz07g15040	C
Photosystem II Subunit P-Like	*PSBPL*	Cz02g18040	C
Photosystem II Subunit Q	*PSBQ1*	Cz05g13170	C
	PSBQ2	Cz05g09100	C
Photosystem II Subunit R	*PSBR*	Cz09g21140	C
Photosystem II Subunit T	*psbT*	CzCPg00410	CE
Photosystem II Subunit W	*PSBW*	Cz19g05110	C
Photosystem II Subunit X	*PSBX*	Cz17g16010	C
Photosystem II Subunit Y	*PSBY*	Cz17g15140	C
Photosystem II Subunit Z	*psbZ*	CzCPg01380	CE
Photosystem II Subunit 27	*PSB27A*	Cz04g12110	O
	PSB27B	Cz09g22190	O
	PSB27C	Cz02g41080	M
Photosystem II Subunit 28	*PSB28*	Cz13g07140	C
Photosystem II Subunit 30	*ycf12*	CzCPg00290	CE
Light Harvesting Complexes			

LHC Proteins Putatively Associated with PSI	*LHC1*	Cz01g23030	C
	LHC2	Cz04g04130	C
	LHC3	Cz04g09210	C
	LHC4	Cz09g32070	C
	LHC5	Cz10g10160	C
	LHC6	Cz12g16070	C
	LHC7	Cz15g17110	C
	LHC8	Cz02g14140	C
	LHC9	Cz16g14150	C
	LHC10	Cz08g26140	O
	LHC11	Cz14g01080	C
	LHC12	Cz03g00160	C
	LHC13	Cz08g00190	SP
Minor PSII Antenna Complex CP29	*LHCB4*	Cz03g26240	O
Minor PSII Antenna Complex CP26	*LHCB5*	Cz08g11160	C
LHC Proteins Putatively Associated with PSII	*LHC14*	Cz04g24050	O
	LHC15	Cz11g23240	C
	LHC16	Cz13g03240	C
	LHC17	Cz07g02110	C
	LHC18	Cz07g02100	C
	LHC19	Cz15g09060	C
	LHC20	UNPLg00080	C
ELIP and HLIP Proteins			
Early Light-Induced Protein Homolog	*ELIP1*	Cz07g29280	C
	ELIP2	Cz07g30110	M
	ELIP3	Cz14g24190	C
	ELIP4	Cz01g42110	C
	ELIP5	Cz02g38160	M
	ELIP6	Cz12g27130	C
	ELIP7	Cz17g00090	C
	ELIP8	Cz17g16050	O
	ELIP9	Cz03g10240	M
	ELIP10	Cz08g00120	C
High Light-Induced Protein Homolog	*HLIP1*	Cz14g04020	C
Chloroplastic ATP Synthase			
CF1 Alpha Subunit	*atpA*	CzCPg01290	CE
CF1 Beta Subunit	*atpB*	CzCPg01090	CE
CF1 Gamma Subunit	*ATPC1*	Cz09g21130	C

	ATPC2	Cz08g05010	C
CF1 Delta Subunit	*ATPD*	Cz05g34180	C
CF1 Epsilon Subunit	*atpE*	CzCPg00310	CE
CF0 B Subunit	*atpF*	CzCPg00020	CE
CF0 B' Subunit	*ATPG*	Cz02g27060	C
CF0 C Subunit	*atpH*	CzCPg00030	CE
CF0 A Subunit	*atpI*	CzCPg00180	CE
Chloroplast Translocase Complexes			
Chloroplast Inner Membrane Translocase	*TIC20A*	Cz06g20080	C
	TIC20B	Cz12g08220	O
	TIC21A	Cz08g26130	M
	TIC21B	Cz16g11150	O
	TIC22	Cz07g31050	O
	TIC40A	Cz02g19150	M
	TIC40B	Cz02g17170	
	TIC110	Cz11g07180	C
Chloroplast Outer Membrane Translocase	*TOC34*	Cz09g17330	O
	TOC64A	Cz06g09170	
	TOC64B	Cz07g13240	
	TOC75A	Cz01g30030	C
	TOC75B	Cz04g11310	
	TOC120	Cz03g17170	O
SRP-Independent Translocase Complex	*TATA*	UNPLg00323	M
	TATB	Cz09g27050	SP
	TATC	Cz10g05030	C
Chloroplast Signal Recognition Particle 43 kDa Subunit	*CPSRP43*	Cz10g07010	SP
Chloroplast Signal Recognition Particle Receptor	*FTSY*	Cz12g26150	C
Signal Recognition Particle Receptor, Alpha Subunit	*CPSRP1*	Cz05g34010	SP
Chloroplast Signal Recognition Particle 54 kDa Subunit	*CPSRP54A*	Cz02g05130	O
	CPSRP54B	Cz08g20150	M
Preprotein Translocase SecY	*SECY1*	Cz04g37160	SP
	SECY2	Cz05g25080	SP
Alpha Subunit of ER Translocon	*SEC61A*	Cz01g29020	SP
Photosystem & Chloroplast Assembly Factor*s*			
PSI Assembly Factor	*ycf3*	CzCPg01160	CE
PSI Assembly Factor	*ycf4*	CzCPg01170	CE
Ycf3-Interacting Protein	*Y3IP1*	Cz15g08190	O
PSII Assembly Factor	*HCF136*	Cz10g21040	O
Low PSII Accumulating	*LPA1A*	Cz05g07170	C

	LPA1B	Cz01g20260	C
	LPA3A	Cz16g08170	C
	LPA3B	Cz12g08020	O
Chloroplast Ser-Thr Kinase 7	*STN7*	Cz06g00060	C
Chloroplast Ser-Thr Kinase 8	*STN8*	Cz10g22110	M
Accumulation/Replication of Chloroplasts	*ARC6A*	Cz06g19140	O
	ARC6B	Cz09g17290	C
	ARC6C	Cz01g24010	O
Potassium/Proton Antiporter	*KEA1*	Cz13g09190	O
	KEA3	Cz07g15180	O
	KEA4	Cz10g20050	O
ATP-Dependent Zn Metalloproteinase	*ftsH*	CzCPg00050	CE
	FTSH	Cz02g25060	C
Chloroplast Rubredoxin	*RBD1*	Cz05g36060	O
PsbP Domain Containing Proteins	*PPD1*	Cz03g13190	C
	PPD2	Cz19g02070	O
	PPD3	Cz14g21260	O
	PPD4	Cz01g04150	O
	PPD5	Cz12g14170	C
	PPD6	Cz04g08120	C
	PPD7	Cz04g12120	C
	PPD8	Cz10g23090	O
	PPD9	Cz04g02050	C
High Chl Fluorescence 101: 4Fe-4S Scaffold	*HCF101*	Cz04g32150	SP
High Chl Fluoresence 164: Trx-Like	*HCF164A*	Cz07g07160	O
	HCF164B	Cz08g24090	C
Membrane Insertion Protein	*OxaA1*	Cz03g01150	M
	OxaA2	Cz01g09150	M
Albino-3	*ALB3A*	Cz05g17260	SP
	ALB3B	Cz05g36130	O
Curvature of the Thylakoid Protein	*CURT1*	Cz05g04030	C
Thylakoid Formation Factor	*THF1*	Cz05g05190	C
Putative Retrograde Signaling Factor, DUF3506 Containing	*EX1*	Cz01g19060	C
	EX2	Cz06g34170	O
PsbA Maturation Factor	*PAM68*	Cz02g08100	C
Chloroplast Sec14-Like Protein	*CPSFL1*	Cz03g12090	C
Proton Gradient Regulator	*PGR7*	Cz11g03230	SP
Fatty Acid Biosynthesis			
Acetyl Co-A Synthase	*ACS1*	Cz12g10100	O
	ACS2	Cz09g15060	O

Acyl-Carrier Protein (ACP)	*ACP1*	Cz07g17120	C
	ACP2	Cz09g30220	C
Acetyl-CoA Carboxylase Carboxyltransferase, beta Subunit	*CAC1*	Cz02g17060	O
Acetyl-CoA Carboxylase Carboxyltransferase, alpha Subunit	*CAC2*	Cz02g12030	O
Acetyl-CoA Carboxylase	*ACC1*	Cz19g10190	O
Biotin Acetyl-CoA Ligase	*BPL*	Cz14g10290	O
Malonyl-CoA:ACP Acyltransferase	*MCT1*	Cz13g05150	O
	MCT2	Cz04g37050	O
Beta-Ketoacyl ACP Synthase	*KAS1*	Cz02g14160	C
	KAS2	UNPLg00257	C
	KAS3	Cz18g03070	O
Beta-Ketoacyl Synthase	*KASX*	Cz06g14030	M
Beta-Ketoacyl ACP Reductase	*KAR1*	Cz01g34370	M
	KAR2	Cz16g00050	O
	KAR3	Cz15g00050	O
	KAR4	Cz04g17270	O
	KAR5	Cz11g27250	SP
Beta-Hydroxyacyl ACP Dehydrase/Dehydratase	*FABZ*	Cz01g09160	C
Enoyl ACP Reductase	*ENR*	Cz11g20040	C
Long-Chain Beta-Ketoacyl Synthase	*LCKAS1*	Cz06g36280	SP
	LCKAS2	Cz12g02090	O
	LCKAS3	Cz12g02100	O
	LCKAS4	Cz13g05060	SP
	LCKAS5	Cz17g10060	SP
	LCKAS6	Cz05g07110	SP
	LCKAS7	Cz11g13060	O
	LCKAS8	Cz19g03050	SP
Fatty Acid Metabolism			
Enoyl-CoA ACP Hydratase	*ECH1*	Cz08g18230	O
	ECH2	Cz11g22170	O
	ECH3	Cz03g36260	O
Enoyl-CoA ACP Hydratase/Isomerase D	*ECHID*	Cz04g19010	O
Acetyl-CoA Acyltransferase Thiolase	*ATO2A*	Cz04g25080	SP
	ATO2B	Cz06g36270	M
Lipid Biosynthesis			
Stearoyl ACP Desaturase	*SACPD1*	Cz04g09090	O
	SACPD2	Cz13g17200	C
Glycerol-3-Phosphate O-Acyltransferase	*GPAT1*	Cz11g03260	C

	GPAT2	Cz09g31330	M
Lysophosphatidylcholine Acyltransferase	*LPCAT*	Cz02g32040	O
Phosphatidylcholine Sterol Acyltransferase	*LCAT1*	Cz17g15210	O
	LCAT2	Cz14g23330	O
Cytidinediphosphate Diacylglycerol Synthase	*PCT1*	Cz01g36190	O
	PCT2	Cz10g20080	SP
Sulfolipid Synthase	*SQD2*	Cz07g23140	O
UPD-Sulfoquinovosyl Synthase	*SQD1*	Cz03g31030	C
Phospholipid : Diacylglycerol Acyltransferase	*PDAT*	Cz10g07210	O
Glycerol-3-Phosphate Dehydrogenase	*GPDH1*	Cz10g29180	O
	GPDH2	Cz12g24180	O
Diacylglycerol Kinase	*DGK1*	Cz05g16250	O
	DGK2	Cz01g12160	O
Diacylglycerol Acyltransferase Type 1	*DGAT1A*	Cz06g04190	SP
	DGAT1B	Cz09g23020	O
	DGAT1C	Cz03g14080	O
Membrane Bound Diacylglycerol Acyltransferase MBOAT, Type 1	*DGAT1D*	Cz09g08290	O
Diacylglycerol Acyltransferase Type 2	*DGAT2A*	Cz08g14220	O
	DGAT2B	Cz11g21100	O
	DGAT2C	Cz11g24150	SP
	DGAT2D	Cz09g27290	SP
	DGAT2E	Cz15g22140	O
Diacylglycerol Acyltransferase Type 2B	*DGAT2F*	Cz06g35060	O
	DGAT2G	Cz06g22030	O
Triacylglycerol Lipase	*LIP1*	Cz02g24260	SP
	LIP2	Cz13g03050	O
Phosphatidylglycerophosphate Synthase	*PGP1*	Cz01g26180	C
	PGP2	Cz06g21300	M
Dihydroxyacetone Kinase	*DAK*	Cz10g02240	O
Triosephosphate Isomerase	*TPI*	Cz06g17270	O
Glycerol-3-Phosphate Dehydrogenase	*GPD1*	Cz10g29180	O
	GPD2	Cz12g24180	O
Fatty Acid-Desaturase A	*FADA*	Cz12g10230	SP
Fatty Acid-Desaturase, omega-3	*FAD7A*	Cz04g31180	C
	FAD7B	Cz06g28130	C
Fatty Acid Desaturase, omega-6	*FAD2A*	Cz03g33220	O
	FAD2B	Cz08g04110	C
Fatty Acid Desaturase, delta-6	*FAD3A*	UNPLg00012	O
	FAD3B	Cz06g12050	C
Fatty Acid Desaturase, delta-9	*FAD5A1*	Cz07g00120	SP

	FAD5A2	Cz06g00170	O
	FAD5C	Cz13g01140	O
Fatty Acid Desaturase, delta-12	*FAD6A*	Cz08g21150	SP
	FAD6B	Cz11g21120	O
Long Chain Acyl-CoA Synthetase	*LACS1*	Cz07g22230	O
	LACS2	Cz11g20120	C
	LACS3	Cz05g30060	O
	LACS4	Cz01g36150	O
Phosphatidate Phosphatase	*PAP*	Cz08g10040	C
CDP:DAG Synthetase	*CDS*	Cz01g36190	O
S-Adenosylmethionine Synthetase	*METM*	Cz15g18200	O
	METK	Cz05g24030	O
Betaine Lipid Synthase	*BTA*	Cz01g13260	O
CDP-Ethanolamine: DAG-Ethanolamine Phosphotransferase	*EPT*	Cz05g09130	O
Serine Decarboxylase	*SDC*	Cz06g24140	O
Ethanolamine Kinase	*ETK*	Cz11g15030	SP
CTP:Phosphoethanolamine Cytidylyltransferase	*ECT*	Cz05g17180	SP
Inositol-3-Phosophate Synthase	*IPS*	Cz01g18130	O
CDP-DAG:Inositol Phosphatidyltransferase	*PIS*	Cz17g13240	O
Monogalactosyldiacylglycerol Synthase	*MGD*	Cz08g30040	O
Digalactosyldiacylglycerol Synthase	*DGD1*	Cz13g19030	O
	DGD2	Cz10g17090	O
Calvin-Benson-Bassham Cycle			
Rubisco Accumulation Factor	*RAF*	Cz09g07180	C
Rubisco Activase	*RCA1*	Cz03g15230	C
	RCA2	Cz13g02230	O
Rubisco Small Subunit	*RBCS*	Cz17g13100	C
Rubisco Large Subunit	*rbcL*	CzCPg00360	C
Rubisco Large Subunit Methyltransferase	*RBCMT*	Cz05g09240	O
Phosphoglycerate Dehydrogenase	*PGD*	Cz01g24020	O
Phosphoglycerate Kinase	*PGK*	Cz16g19260	C
Glyceraldehyde-3-Phosphate Dehydrogenase	*GAPDH*	Cz05g34160	C
Triose Phosphate Isomerase	*TPI*	Cz06g17270	O
Fructose-Bisphosphate Aldolase	*FBA1*	Cz03g13070	O
	FBA2	Cz06g07090	O
	FBA3	Cz05g37140	C
Fructose-2,6-Bisphosphate-2-Phosphatase	*FBP3A*	Cz16g19090	O
	FBP3B	Cz12g10060	O
Fructose-1,6-Bisphosphatase	*FBP1*	Cz05g01180	C

	FBP2	Cz04g03070	C
Transketolase	*TRK*	Cz03g04080	C
Sedoheptulose-1,7-Bisphosphatase	*SBP*	Cz01g42020	C
Ribulose-5-Phosphate-3-Epimerase	*RPE2*	Cz04g31230	O
	RPE1	Cz05g11190	C
Ribose-5-Phosphate Isomerase	*RPI1*	Cz09g17220	C
	RPI2	Cz01g08190	M
Phosphoribulokinase	*PRK*	Cz07g08080	O
TCA / Glyoxylate Cycle			
Citrate Synthase	*CIT*	Cz02g27080	M
Aconitase	*ACH1*	Cz13g00140	C
Isocitrate Lyase	*ICL*	Cz07g04220	O
Malate Synthase	*MAS1*	Cz14g25060	O
Malate Dehydrogenase	*MDH1*	Cz01g16230	C
	MDH2	Cz02g21340	O
	MDH3	UNPLg00180	O
	MDH4	Cz12g06020	C
	MDH5	Cz04g33150	C
Isocitrate Dehydrogease	*IDH1*	Cz11g08120	M
	IDH2	Cz11g28180	M
2-Oxoglutarate Dehydrogenase, E1 Subunit	*OGD*	Cz05g03220	M
Succinyl-CoA Synthetase/Ligase, alpha	*SCL1*	Cz01g33150	O
Succinyl-CoA Synthetase/Ligase, beta	*SCL2*	Cz03g31200	O
Succinate Semialdehyde Dehydrogenase	*SSADH*	Cz05g02090	M
Succinate Dehydrogenase Subunit 1	*SDH1*	Cz03g18320	M
Succinate Dehydrogenase Subunit 2	*SDH2*	Cz15g13230	M
Succinate Deydrogenase, b560 Subunit	*SDH3*	Cz07g14020	M
Succinate Dehydrogenase Subunit 4	*SDH4*	Cz07g14015	O
Fumarase	*FUM*	Cz07g04030	SP
Glycolysis			
Sucrose Synthase	*SUC1*	Cz05g24180	M
Hexokinase	*HXK1*	Cz13g07170	O
UDP-Glucose Pyrophosphorylase	*UGP*	UNPLg00641	O
	UGP2	Cz07g13190	C
Phosphoglucomutase	*GPM1*	Cz04g03150	C
	GPM2	Cz15g21100	O
Phosphofructokinase	*PFK1*	Cz09g25120	C
	PFK2	Cz07g13120	C

Glucose-6-Phosphate Dehydrogenase	*GPDH1*	Cz03g12030	C
	GPDH2	Cz06g12080	C
6-Phosphogluconate Dehydrogenase	*GND*	Cz05g06160	M
Pyruvate Carboxylase	*PYC*	Cz04g02090	SP
PEP Carboxykinase	*PCK*	Cz01g05160	C
PEP Carboxylase	*PEPC1*	UNPLg00263	O
	PEPC2	Cz01g21210	O
	PEPC3	Cz14g19100	C
Phosphoglycerate Mutase	*PGM1*	Cz04g11130	O
	PGM3	Cz14g01200	C
	PGM4	Cz14g08090	C
	PGM5	Cz06g01110	O
	PGM7	Cz01g38050	M
Enolase	*ENO*	Cz05g10010	O
Pyruvate Kinase	*PYK1*	Cz15g09100	O
	PYK2	Cz01g21060	O
	PYK3	Cz14g14130	O
	PYK4	Cz10g06190	O
	PYK5	Cz16g00040	O
Pyruvate Decarboxylase	*PDC*	Cz01g33100	O
NADP-Malic Enzyme	*MME1*	Cz15g01060	M
	MME2	Cz15g18140	O
	MME4	Cz09g11240	O
	MME5	Cz07g26110	O
	MME6	Cz04g02110	SP

Table S4. ***Chromochloris zofingiensis*** **astaxanthin-deficient mutant characteristics.**

Mutant	Phenotype	Mutation location	Region	Molecular basis of the mutation
bkt1-1	no astaxanthin	164 aa	highly conserved	Pro to Ser
bkt1-2	no astaxanthin	275 aa	highly conserved	Pro to Leu
bkt1-3	no astaxanthin	163 aa	highly conserved	Asp to Asn
bkt1-4	no astaxanthin	225 aa	medium conserved	Arg to stop
bkt1-6	no astaxanthin	210 aa	medium conserved	Leu to stop
bkt1-7	no astaxanthin	131 aa	highly conserved	Gly to Asp
bkt1-9	no astaxanthin	165 aa	highly conserved	Asp to Val
bkt1-14	no astaxanthin	251 aa	highly conserved	Ser to Leu
bkt1-15	no astaxanthin	165 aa	highly conserved	Pro to Thr
bkt1-16	no astaxanthin	166 aa	highly conserved	Asp to Glu
bkt1-17	no astaxanthin	31 aa	early stop	Gln to Stop
bkt1-18	no astaxanthin	99 aa	highly conserved	Leu to Pro
bkt1-21	no astaxanthin	79 aa	early stop	Trp to Met (2 mutations) followed by a deletion of 5 base frameshift to a stop after 7 aa
bkt1-22	no astaxanthin	158 aa	highly conserved	Thr to Ile
bkt1-24	no astaxanthin	120 aa	medium conserved	Ile to Phe
bkt1-25	no astaxanthin	1st intron	early deletion or partially deletion	314bp deletion from 1st intron into second exon

Table S5. PCR primers used to sequence beta-ketolase in *Chromochloris zofingiensis.*

Gene	Forward Primer (5'-3')	Reverse Primer (5'-3')
BKT1	TACTCAGGCATCTACGTGTT	TGCGAACAACTCAAAGCATA
BKT2	ATTCAGGCGACTACATGACTGG	GTTGCATGGCTTTCTCACATCATT

Datasets key (for Datasets S1–S21)

Suggested filenames are for consistency with the project's website: http://genomes.mcdb.ucla.edu/Chromochloris/ .

Assembly

Genome nucleotide sequences in plain text FASTA format ("ChrZof version 5 release of 2017-01-01"):

Dataset S1 (ChrZofV5.lcMaskByRepeatMasker-on-RepeatModeler_plus_AllRepBase20160829.fasta) — both;
Dataset S2 (ChrZofV5.lcMaskByRepeatMasker-on-RepeatModeler.fasta) — masked via RepeatModeler;
Dataset S3 (ChrZofV5.lcMaskByRepeatMasker-on-AllRepBase20160829.fasta) — masked via RepBase:
The three flavors differ in their level of lowercase repeat masking. Dataset S1 is suggested as a default.

Known assembly issues for ChrZofV5 in Microsoft Excel format:

Dataset S4 (ChrZofV5-KnownAssemblyIssues.xls) — list;
Dataset S5 (ChrZofV5-KnownAssemblyIssues-Summary.xls) — summary of list.

Optical map (of nominal date 2017-01-01):

Dataset S6 (ChrZof-OpticalMap-asOrderedListOfBamHIopticalFragmentLengthsInBP.txt) — plain text giving, for each nuclear chromosome, estimated lengths of successive BamHI complete digest fragments in nucleotide basepairs, traveling 5′ to 3′ along plus strands.
Dataset S7 (ChrZof-OpticalMap-in-OpGen-MapSolver-XML-Format.xml) — OpGen MapSolver XML version.

(Project website file 'RepeatMaskerRepeatModeler-ChrZof.tar' with details of repeat analysis was too complex to include with the journal publication and is omitted as a Dataset.)

tRNAs

Dataset S8 (ChrZofV5.lcMaskByRepeatMasker-on-RepeatModeler_plus_AllRepBase20160829.noCpMt.tRNAscan-SE-1.3.1.output.txt) — plain text direct from tRNAscan-SE 1.3.1.

Gene families

Dataset S9 (ChrZof-v5.2-OrthologsV3.famId-ChrZofs-ChlReis-AraThas-CocSubs-Chlores-MonNegs-numPrimaryGenesSameOrgoOrder-numAdditionalGenesSameOrgoOrder.txt) —
plain text with tab-separated columns giving 3rd version of orthologs based on *C. zof.* version 5.2 gene models:
Column 1: family accession in pattern "`CzOrth_5.2-3_#####`";
Columns 2, …, 7: comma-separated gene names for Table S1 genome *Chromochloris zofingiensis*, *Chlamydomonas reinhardtii*, *Arabidopsis thaliana*, *Coccomyxa subellipsoidea* C-169, *Chlorella sp.* NC64A, and *Monoraphidium neglectum*, respectively ('`&`' suffix to gene name indicates additional gene, else primary);
Columns 8, …, 13: for same order of organisms, number of primary genes as non-negative decimal integer;
Columns 14, …, 19: for same order of organisms, number of additional genes as non-negative decimal integer.

Dataset S10 (InfoExtractsOrthosV3-CzHasOneOrMoreOf520handTags.keepLongInfos.txt) — ≥1 *C. zof.* hand-tag;
Dataset S11 (InfoExtractsOrthosV3-CzHasOneOrMoreOf520handTags.omitLongInfos.txt) — ≥1 *C. zof.* hand-tag;
Dataset S12 (InfoExtractsOrthosV3-CzNoneHandTagged.keepLongInfos.txt) — ≥1 *C. zof.* but none hand-tagged;
Dataset S13 (InfoExtractsOrthosV3-CzNoneHandTagged.omitLongInfos.txt) — ≥1 *C. zof.* but none hand-tagged;
Dataset S14 (InfoExtractsOrthosV3-PureNonCZ.keepLongInfos.txt) — families with no *C. zof.* genes;
Dataset S15 (InfoExtractsOrthosV3-PureNonCZ.omitLongInfos.txt) — families with no *C. zof.* genes:
These plain text files give compilations of various useful pieces of information from genome projects of Table S1 for the ChrZof-v5.2-OrthologsV3 gene families. Datasets S10/S12/S14 have all information (including long free text), while S11/S13/S15 keep only shorter information. Families partition to S10/S11 vs. S12/S13 vs. S14/S15 according to if they involve at least one hand-tagged *C. zof.* gene, else any *C. zof.* gene, and all the rest.

Dataset S16 (ChrZof-v5.2-OrthologsV3.geneSymbols-AraTha-ChlRei-ChrZof.ChrZofYesHandSymboled.txt);
Dataset S17 (ChrZof-v5.2-OrthologsV3.geneSymbols-AraTha-ChlRei-ChrZof.ChrZofNotHandSymboled.txt);
Dataset S18 (ChrZof-v5.2-OrthologsV3.geneSymbols-AraTha-ChlRei-ChrZof.ChrZofNotHaveAnyGenes.txt):
These tab-separated plain text files compile gene symbol annotations from genome projects of Table S1 for the ChrZof-v5.2-OrthologsV3 gene families:
Column 1: family accession in pattern "`CzOrth_5.2-3_#####`";
Columns 2, …, 4: gene symbols associated to the *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, and *Chromochloris zofingiensis* genes, respectively, in the gene family.
Families are in Dataset S16 if have ≥1 *C. zof.* hand-symboled gene, else S17 if ≥1 *C. zof.* gene, else S18.

Genes

Dataset S19 (ChrZof.annotations.v5.2.3.2.gff3) — GFF3-format ver. 5.2.3.2 gene models on ChrZofV5 assembly.

Expression

Dataset S20 (transcriptome_RNA-Seq_FPKMs.txt) — plain text matrix with rows for *C. zof.* genes (with identifier in first column) and columns for FPKMs over diverse series conditions.

Dataset S21 (high-light_RNA-Seq_FPKMs.txt) — plain text matrix with rows for *C. zof.* genes (with identifier in first column) and columns for FPKMs over high light series conditions.

SI References

1. Parkinson DY*, et al.* (2013) Nanoimaging cells using soft X-ray tomography. *Methods Mol Biol* 950:457-481.
2. Le Gros MA*, et al.* (2012) Visualizing sub-cellular organization using soft X-ray tomography. *Comprehensive Biophysics,* Biophysical Techniques for Characterization of Cells, ed Egelman EH (Academic Press, Oxford), Vol 2, pp 90-110.
3. Le Gros M*, et al.* (2014) Biological soft X-ray tomography on beamline 2.1 at the Advanced Light Source. *J Synchrotron Radiat* 21(Pt 6):1370-1377.
4. Kremer JR, Mastronarde DN, & McIntosh JR (1996) Computer visualization of three-dimensional image data using IMOD. *J Struct Biol* 116(1):71-76.
5. https://github.com/vsbuffalo/scythe
6. Dobin A*, et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.
7. Trapnell C*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562-578.
8. Goff L, Trapnell C, & Kelley D (2013) cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. *R package version 2.16.0.*
9. Gautier L, Cope L, Bolstad BM, & Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3):307-315.
10. Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550.
11. Boisvert S, Raymond F, Godzaridis E, Laviolette F, & Corbeil J (2012) Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 13(12):R122.
12. Simpson JT*, et al.* (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19(6):1117-1123.
13. Ribeiro FJ*, et al.* (2012) Finished bacterial genomes from shotgun sequence data. *Genome Res* 22(11):2270-2277.
14. Gnerre S*, et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108(4):1513-1518.
15. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
16. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth* 9(4):357-359.
17. Kim D, Langmead B, & Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Meth* 12(4):357-360.
18. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403-410.
19. Camacho C*, et al.* (2009) BLAST+: architecture and applications. *BMC Bioinf* 10:421.
20. Kielbasa SM, Wan R, Sato K, Horton P, & Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res* 21(3):487-493.
21. Harris R (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis (The Pennsylvania State University).
22. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12(4):656-664.

23. Chaisson MJ & Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinf* 13:238.
24. Daily J (2016) Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinf* 17:81.
25. Hackl T, Hedrich R, Schultz J, & Förster F (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30(21):3004-3011.
26. https://github.com/jstjohn/SeqPrep
27. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10-12.
28. Marçais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-770.
29. Kurtz S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12.
30. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl Acids Res* 27(2):573-580.
31. Warburton PE, Giordano J, Cheung F, Gelfand Y, & Benson G (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* 14(10a):1861-1869.
32. Smit A, Hubley R, & Green P (2013-2015) RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
33. Bao W, Kojima KK, & Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11.
34. Smit A & Hubley R (2008-2015) RepeatModeler Open-1.0. <http://www.repeatmasker.org>.
35. Stanke M, Schoffmann O, Morgenstern B, & Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinf* 7:62.
36. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl Acids Res* 25(5):0955-0964.
37. Fiume M, Williams V, Brook A, & Brudno M (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics* 26(16):1938-1944.
38. Freese NH, Norris DC, & Loraine AE (2016) Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* 32(14):2089-2095.
39. Thorvaldsdóttir H, Robinson JT, & Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14(2):178-192.
40. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.
41. Li H*, et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
42. https://github.com/thegenemyers/DEXTRACTOR
43. Anders S, Pyl PT, & Huber W (2014) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166-169.

44. Rice P, Longden I, & Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276-277.
45. https://www.ncbi.nlm.nih.gov/
46. Finn RD*, et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucl Acids Res* 44(D1):D279-D285.
47. Nawrocki EP*, et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucl Acids Res* 43(D1):D130-D137.
48. Lagesen K*, et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35(9):3100-3108.
49. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814):796-815.
50. Haas BJ*, et al.* (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucl Acids Res* 31(19):5654-5666.
51. Blanc G*, et al.* (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol* 13(5):R39.
52. Merchant SS*, et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318(5848):245-251.
53. Blanc G*, et al.* (2010) The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* 22(9):2943-2955.
54. Bogen C*, et al.* (2013) Reconstruction of the lipid metabolism for the microalga *Monoraphidium neglectum* from its genome sequence reveals characteristics suitable for biofuel production. *BMC Genomics* 14:926.
55. Chandrasekhara C, Mohannath G, Blevins T, Pontvianne F, & Pikaard CS (2016) Chromosome-specific NOR inactivation explains selective rRNA gene silencing and dosage control in *Arabidopsis*. *Genes Dev* 30(2):177-190.
56. Huerta-Cepas J, Serra F, & Bork P (2016) ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 33(6):1635-1638.
57. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3):e9490.
58. Leliaert F*, et al.* (2012) Phylogeny and molecular evolution of the green algae. *Crit Rev Plant Sci* 31(1):1-46.
59. https://www.blast2go.com/
60. Consortium TGO (2015) Gene Ontology Consortium: going forward. *Nucl Acids Res* 43(D1):D1049-D1056.
61. Eddy SR (2012) A new generation of homology search tools based on probalistic inference. *Genome Informatics 2009*, (Imperial College Press), pp 205-211.
62. http://geneontology.org/external2go/pfam2go (2016/09/17 11:36:45).
63. Baroli I, Do AD, Yamane T, & Niyogi KK (2003) Zeaxanthin accumulation in the absence of a functional xanthophyll cycle protects *Chlamydomonas reinhardtii* from photooxidative stress. *Plant Cell* 15(4):992-1008.