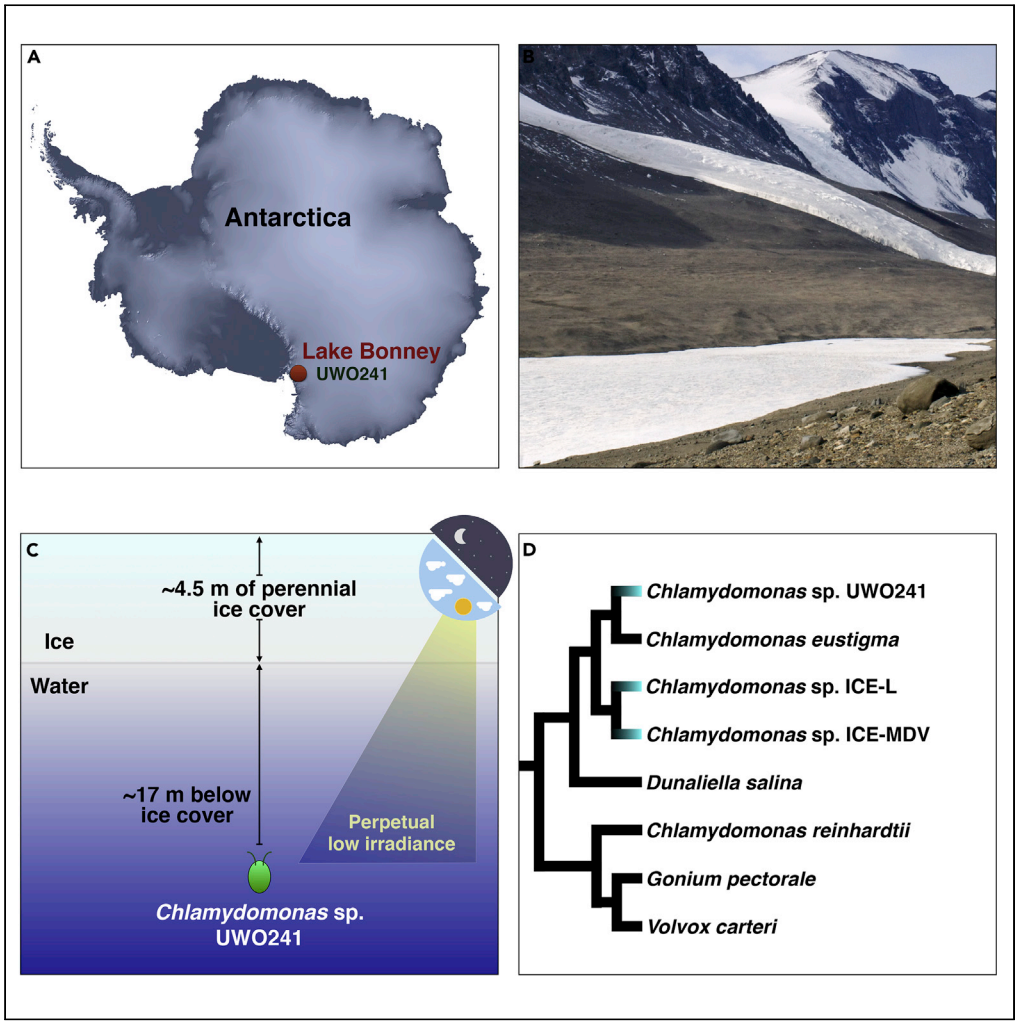
iScience

CellPress
OPEN ACCESS

Article

Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241



Xi Zhang, Marina Cvetkovska, Rachael Morgan-Kiss, Norman P.A. Hüner, David Roy Smith

nhuner@uwo.ca (N.P.A.H.)
dsmit242@uwo.ca (D.R.S.)

HIGHLIGHTS

Chlamydomonas sp. UWO241 is a green alga originating from Lake Bonney, Antarctica

We present a draft nuclear genome sequence of UWO241 (~212 Mb).

The UWO genome contains hundreds of highly similar duplicated genes

These duplicates, we argue, might be involved in cold adaptation

Zhang et al., iScience *24*, 102084
February 19, 2021
https://doi.org/10.1016/j.isci.2021.102084

iScience

CellPress
OPEN ACCESS

Article

Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241

Xi Zhang,[1] Marina Cvetkovska,[2] Rachael Morgan-Kiss,[3] Norman P.A. Hüner,[1,*] and David Roy Smith[1,4,*]

SUMMARY

Antarctica is home to an assortment of psychrophilic algae, which have evolved various survival strategies for coping with their frigid environments. Here, we explore Antarctic psychrophily by examining the ~212 Mb draft nuclear genome of the green alga *Chlamydomonas* sp. UWO241, which resides within the water column of a perennially ice-covered, hypersaline lake. Like certain other Antarctic algae, UWO241 encodes a large number (≥37) of ice-binding proteins, putatively originating from horizontal gene transfer. Even more striking, UWO241 harbors hundreds of highly similar duplicated genes involved in diverse cellular processes, some of which we argue are aiding its survival in the Antarctic via gene dosage. Gene and partial gene duplication appear to be an ongoing phenomenon within UWO241, one which might be mediated by retrotransposons. Ultimately, we consider how such a process could be associated with adaptation to extreme environments but explore potential non-adaptive hypotheses as well.

INTRODUCTION

"Antarctica is a very alien environment, and you can't survive here more than minutes if you're not equipped properly and doing the right thing all the time" — Jon Krakauer

Life persists in the harshest places. Take *Chlamydomonas* sp. UWO241, for instance. This unicellular chlorophycean green alga—formerly called *Chlamydomonas raudensis* (Possmayer et al., 2016)—resides in the permanently ice-covered Lake Bonney (McMurdo Dry Valleys, Antarctica) (Pocock et al., 2004) (Figures 1A and 1B) and is so attuned to cold water (~5°C year-round) that it cannot grow above 18°C, making it a true psychrophile (Morita, 1975). Moreover, its position within the water column (~17 m below surface) brings other challenges: high salinity (0.7 M), low levels of phosphorus (N:P ~1000), seasonal extremes in photoperiod, unusually high oxygen concentrations (200% air saturation), and perpetually low irradiance (Morgan-Kiss et al., 2006; Dore and Priscu, 2001) (Figure 1C). How has UWO241 adapted to this extreme habitat?

Recently, it was shown that UWO241, unlike other surveyed eukaryotic algae, produces two near-identical copies of photosynthetic ferredoxin (PETF), resulting from a duplication of the nuclear *petf* gene (Cvetkovska et al., 2018). The retention and expression of this duplicate gene is hypothesized to be an adaptation to the cold, leading to higher protein accumulation (i.e., gene dosage); indeed, UWO241 accumulates greater amounts of PETF than its mesophilic relative *Chlamydomonas reinhardtii* (Merchant et al., 2007; Cvetkovska et al., 2018). Similarly, UWO241 expresses three isoforms of an unusual bidomain enzyme, allowing it to produce high levels of osmoprotectant glycerol (>400 mM) (Kalra et al., 2020). If gene dosage is contributing to psychrophily in UWO241, one might expect other genes to be duplicated. Here, we show that this expectation is true.

Genome sequencing of UWO241 exposed hundreds of gene duplicates for crucial cellular pathways and dozens of genes encoding ice-binding proteins (IBPs). These findings for UWO241 (isolated from a constantly cold but non-freezing environment) mirror many of those from the recent genomic analysis of *Chlamydomonas* sp. ICE-L (Zhang et al., 2020), which originates from a cold but fluctuating Antarctic sea ice environment (Figure 1A), and enhance our understanding of photopsychrophily and the evolutionary dynamics within ice-covered Antarctic lakes.

[1]Department of Biology, University of Western Ontario, London, ON N6A 5B7, Canada

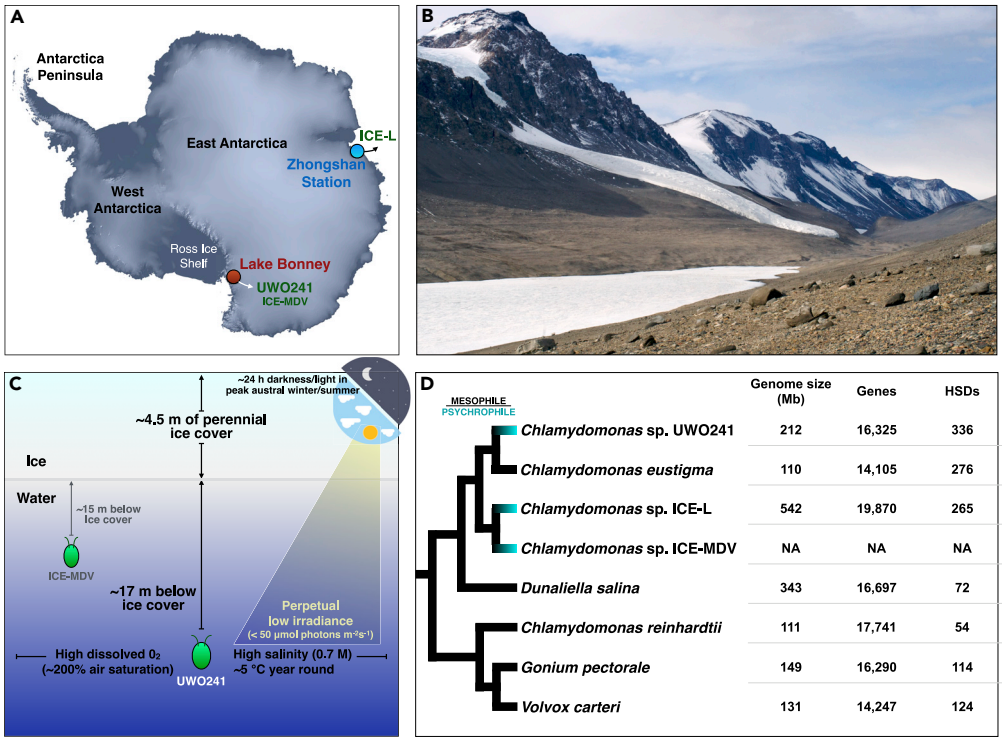
[2]Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada

[3]Department of Microbiology, Miami University, Oxford, OH 45056, USA

[4]Lead contact

*Correspondence: nhuner@uwo.ca (N.P.A.H.), dsmit242@uwo.ca (D.R.S.)

https://doi.org/10.1016/j.isci.2021.102084



	Genome size (Mb)	Genes	HSDs
Chlamydomonas sp. UWO241	212	16,325	336
Chlamydomonas eustigma	110	14,105	276
Chlamydomonas sp. ICE-L	542	19,870	265
Chlamydomonas sp. ICE-MDV	NA	NA	NA
Dunaliella salina	343	16,697	72
Chlamydomonas reinhardtii	111	17,741	54
Gonium pectorale	149	16,290	114
Volvox carteri	131	14,247	124

Figure 1. *Chlamydomonas* sp. UWO241

(A) Origins of isolation of UWO241 and *Chlamydomonas* sp. ICE-MDV (Lake Bonney), as well as *Chlamydomonas* sp. ICE-L (sea ice off of Zhongshan Station); image from NASA Earth Observatory.

(B) Photograph of Lake Bonney (Wikimedia-Commons, 2020).

(C) Simplified diagram showing environmental conditions in Lake Bonney.

(D) Tree of various chlamydomonadalean algae and their nuclear genome statistics; branching order based on previous phylogenetic analyses (Nakada et al., 2008; Possmayer et al., 2016; Zhang et al., 2020).

RESULTS AND DISCUSSION

Draft nuclear genome sequence of an alga from an Antarctic lake

The haploid nuclear genome of UWO241 was assembled *de novo* using a combination of long-read PacBio (~16.5 Gb) and short-read Illumina (~40 Gb) data, resulting in 2,458 scaffolds (N50 = 375.9 kb) with an accumulative length of 211.6 Mb (%GC = 60.6) (Figures 1D and S1). This length is consistent with flow cytometry and *k*-mer spectral analysis of UWO241, which predicted an overall genome size of ~230 Mb (Figure S1). In total, 16,325 protein-coding genes were annotated (all supported by transcriptomic data), capturing ~85% of the Chlorophyte Benchmarking Universal Single-Copy Orthologs data sets (Figure S1), indicating a high level of gene region completeness. The UWO241 genome is rich in functional RNAs (630 tRNAs and 480 rRNAs) as well as noncoding DNA (~87%), having the highest average intron density yet observed from a green alga (~10 introns/gene; avg. intron length 0.9 kb). The intergenic regions abound with repeats, accounting for ~104 Mb (~49%) of the total assembly length, ~70 Mb of which are represented by transposable elements (TEs) (discussed below).

The past decade has brought draft nuclear genomes for >25 different green algal species, with especially strong sampling from the order Chlamydomonadales (Chlorophyceae) (Figure 1D). The UWO241 genome is the second to be sequenced from the Moewusinia clade of the Chlamydomonadales (Nakada et al., 2008), the other coming from the acidophilic species *Chlamydomonas eustigma* (Hirooka et al., 2017). The Moewusinia is closely affiliated with the Monadinia, the clade to which the Antarctic psychrophiles ICE-L and *Chlamydomonas* ICE-MDV belong (Figures 1A and 1D) (Demchenko et al., 2012). Keep in mind that the *Chlamydomonas* genus is polyphyletic and that UWO241, *C. eustigma*, and ICE-L branch closer to *Dunaliella salina* (Figure 1D), for example, than they do to *C. reinhardtii*, which hails from the Reinhardtinia clade (Possmayer et al., 2016; Zhang et al., 2020). What immediately stands out for the UWO241

Table 1. Summary statistics of highly similar duplicate genes (HSDs) in UWO241.

Database	Example identifiers[a]	Number of HSDs (%)[b]	Number of gene copies (%)[b]
Pfam			
Chlorophyll A-B binding protein	PF00504	4 (1%)	25 (2%)
Ribosomal protein	PF01015; PF01775; PF00828	19 (5%)	42 (3%)
Core histone H2A/H2B/H3/H4	PF00125	5 (1%)	99 (7%)
Ice-binding protein (DUF3494)	PF11999	8 (2%)	21 (2%)
Reverse transcriptases	PF00078	38 (11%)	151 (11%)
KEGG			
09,101 Carbohydrate metabolism	K13979 (alcohol dehydrogenase)	12 (4%)	89 (7%)
09,102 Energy metabolism	K02639 (ferredoxin); K08913 (light-harvesting complex II chlorophyll a/b binding protein 2)	10 (3%)	51 (4%)
09,103 Lipid metabolism	K01054 (acylglycerol lipase)	3 (1%)	15 (1%)
09,122 Translation	K02868 (large subunit ribosomal protein L11e)	27 (8%)	47 (4%)
Hypothetical proteins	NA	125 (37%)	357 (27%)

[a]Not all identifiers are listed.
[b]A total of 336 HSDs were identified within the UWO241 genome, encompassing 1,339 gene copies. HSDs share ≥90% pairwise amino acid identity and have lengths within 10 amino acids of each other.

genome as compared to other available green algal nuclear DNAs (nucDNAs) is its relatively large size (approximately twice that of *C. reinhardtii*), record-setting intron density, and high repeat content, outdone only by that of ICE-L (~64% repeats) (Zhang et al., 2020). However, close inspection of the UWO241 coding regions uncovered something much more unique: widespread gene duplication to a degree unmatched in any chlorophyte studied to date.

Hundreds of gene duplicates

Functional annotation of the 16,325 RNA-supported gene models revealed the standard cohort of proteins typically encoded in green algal nuclear genomes (Data S1), as well as many hypothetical proteins (21.8%), paralleling the trends from other available chlamydomonadalean nuclear gene sets, which are generally 20-30% hypothetical. There were no obvious signs of contamination in the assembly or annotations and, with one conspicuous exception (discussed below), little evidence of horizontal gene transfer (HGT). When examining the annotations in detail, it became obvious that many genes were represented two or more times within the assembly. To explore the validity of these multi-copy genes, we performed a series of BLAST (Basic Local Alignment Search Tool) -based analyses with strict downstream filtering.

A protein BLAST of the UWO241 gene models against themselves (E-value cut-off 10^{-5}) detected 901 putative duplicates (encompassing 2,012 gene copies) all with pairwise amino acid identities ≥80%. We filtered this gene set to only those with near-identical protein lengths (within 10 amino acids) and ≥90% pairwise identities, giving a pared-down list of 336 highly similar duplicates (HSDs), totaling 1,339 gene copies (Table 1; Data S2). By setting such a strict cutoff, we have undoubtedly removed some genuine duplicates from this list, but we would rather be conservative in our approach, ensuring that the gene pairs in question are bona fide duplicates rather than spurious ones. The protein sequences of the HSDs were searched against the KEGG (Kyoto Encyclopedia of Genes and Genomes) and Pfam databases, providing a functional breakdown (Table 1; Data S2). HSDs in UWO241 are involved in various cellular pathways, including gene expression, cell growth, membrane transport, and energy metabolism, but also include hypothetical proteins (~37%) and reverse transcriptases (11%) (Table 1; Data S2). HSDs for protein translation, DNA packaging, and photosynthesis were particularly prevalent, with 19 duplications of genes for ribosomal proteins, 10 for histones, and at least 4 for proteins of the chlorophyll a/b binding light-harvesting complex (Table 1; Figures 2A–2C). As with the previously described *petf* duplication (Cvetkovska et al., 2018), many of these HSDs are virtually indistinguishable from each other at the amino acid level and 65 are identical across their nucleotide coding regions (Data S2).

The arrangements of the HSDs are informative. Approximately, 20% contain gene copies that are situated close to one another (often in a head-to-head or head-to-tail orientation) and have very similar intron

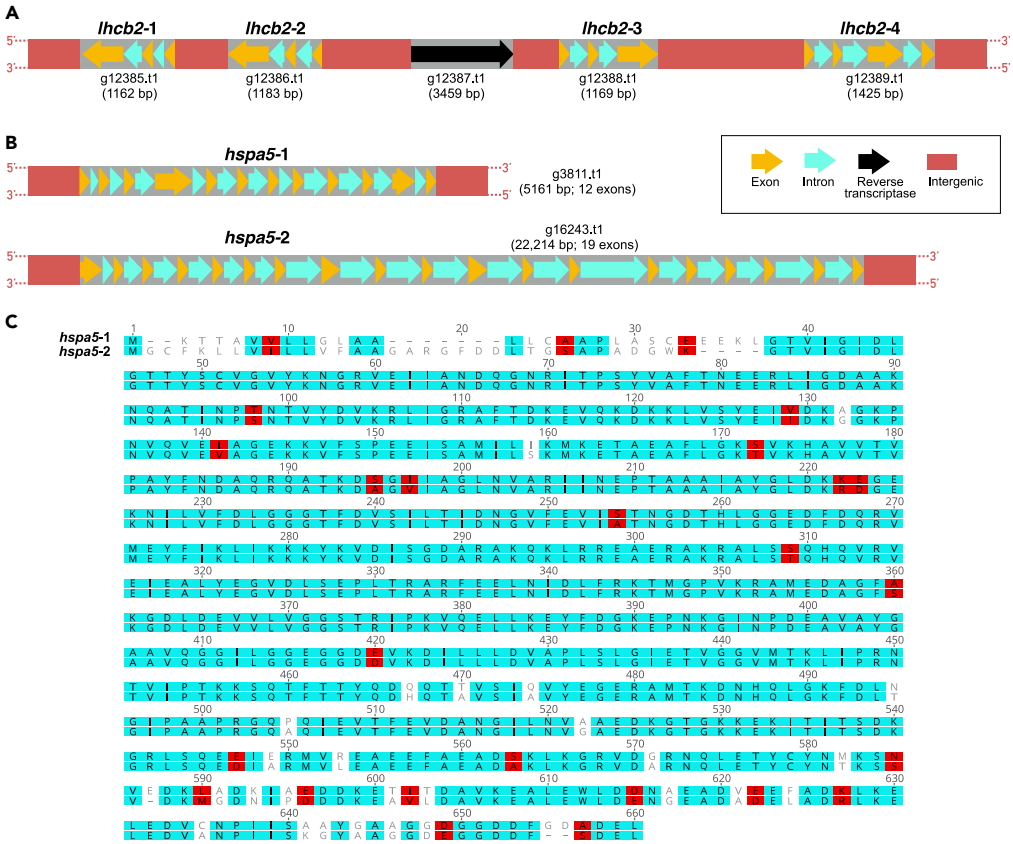


Figure 2. Examples of duplicate genes in *Chlamydomonas* sp. UWO241

(A) Four distinct copies of *lhcb2*, all located on scaffold scf7180000014917.

(B) Two distinct copies of *hspa5*, located on scaffolds scf7180000011611 (*hspa5*-1) and scf7180000015050 (*hspa5*-2).

(C) Pairwise alignment of the deduced amino acid sequences of *hspa5*-1 and *hspa5*-2.

numbers and intronic sequences (based on pairwise alignments), implying that they result from recent tandem duplication events (Figure 2A; Data S2). A clear example of this is the duplication of the *lhcb2* gene (Figure 2A). The remaining HSDs are generally far apart (most on distinct scaffolds) and, despite their matching coding regions, many (~50%) have un-alignable intronic sequences and differing numbers of introns, suggesting that they derive from more ancient duplication events (Figures 2B and 2C; Data S2). This is the case for *petf* (Cvetkovska et al., 2018), as well as for *hspa5* (encoding heat shock 70-kDa protein 5), the two copies of which are found in the middle of distinct scaffolds and share 93% coding sequence identity but <25% nucleotide similarity across their introns (Figures 2B and 2C). Whatever their arrangement, the exonic sequences of more than half of the HSDs (~190) are under strong purifying selection as evidenced by low (<1) nonsynonymous to synonymous substitution rates (dN/dS), ranging from 0 to 0.5 (avg. = 0.2) (Figure S2). It is possible that the strong coding sequence preservation of these duplicates could be aiding the survival of UWO241 in Lake Bonney, perhaps due to increased gene dosage (Innan and Kondrashov, 2010; Kondrashov, 2012), as previously suggested (Cvetkovska et al., 2018). But various non-adaptive explanations are also plausible. The HSDs, however, represent only a fraction of duplicated regions within the genome.

A genome in upheaval

The UWO241 nucDNA contains thousands of partial duplicates, characterized by gene fragments and pseudogenes, as well as duplicated segments of intergenic and intronic DNA (Figure S3; Data S3). These incomplete duplicates, which range in size from ~100-12,000 bp, can exist in high copy numbers (>6) and, like the HSDs, can be found in tandem or on different scaffolds (Figure S3; Data S3). But unlike the HSDs, they are in various states of decay, possibly reflecting an ongoing birth-death process, which is supported

by the fact that many of the complete and partial duplicates are directly associated with or occur near to retrotransposons (RTs) (Figure S3; Data S3), as outlined for the duplication of *Ihcb2* in Figure 2A.

RT-mediated gene duplication is a recurring theme within nuclear genomes (Qian and Zhang, 2014; Panchy et al., 2016; Casola and Betrán, 2017; Kubiak and Makałowska, 2017), including those of green algae (Jąkalski et al., 2016), and the UWO241 genome contains the standard hallmarks of such a phenomenon, such as poly-(A) tail insertions and target site duplications (Figure S3). But this certainly does not rule out the possibility that other processes, such as unequal crossing-over (Zhang, 2003), are contributing to gene duplication within UWO241. Do note that 83% of the HSDs contain introns, a characteristic not generally associated with RT-mediated duplications, but not unprecedented (Casola and Betrán, 2017; Kubiak and Makałowska, 2017). Retrocopies often inherit introns from parental genes, flanking genomic DNA, or the fusion of transcripts (Catania and Lynch, 2008; Zhu et al., 2009; Szcześniak et al., 2011; Kang et al., 2012; Zhang et al., 2014). Altogether, we identified 401 putatively functional RTs in the nucDNA, including 77 long terminal repeat (LTR) and 324 non-LTR RTs. These numbers are likely underestimates of the true total as they do not include retropseudogenes, partial retroelements, or identified RTs with no RNA-seq support, which together account for >10% of the assembly. What's more, there are >480 duplicated regions containing a reverse-transcriptase domain, including ones in noncoding DNA. UWO241 has more retroelements than all other surveyed chlorophytes (4 times that of *C. reinhardtii*) with the exception of ICE-L, for which non-LTR RTs account for a staggering ~23% of the genome (Zhang et al., 2020). In addition to RTs, the UWO241 and ICE-L genomes share another atypical feature—genes for IBPs.

Horizontally acquired and duplicated ice-binding proteins

The UWO241 genome encodes no fewer than 37 proteins with an ice-binding domain (DUF3494) (Figure 3A), which is among the largest number of IBPs ever recorded in a photosynthetic protist. This wealth of IBPs appears to be the consequence of HGT events in combination with gene duplication. Phylogenetic analyses of the IBP genes, which range in size from 483-37,549 bp, show their similarity to type I bacterial and archaeal IBPs (Figure 3B), which is consistent with previous work (Raymond and Morgan-Kiss, 2013). Nuclear genes acquired via recent HGT events from bacteria usually lack introns (Keeling and Palmer, 2008), as do 14 of the IBP genes from UWO241; the remaining genes, with 4 exceptions, all have a single, short intron at their 3′ ends. The largest IBP gene, however, contains 29 introns. The IBP genes show varying degrees of similarity with each other (Figure 3C), including 8 groupings of almost identical genes, suggesting a complicated history of IBP gene acquisition and duplication within UWO241. The presence of pseudogenes and gene fragments with similarity to IBPs (Data S3) indicates that some previously functional IBP coding regions might have been lost.

These findings add to the growing list of psychrophilic and psychrotolerant algae encoding IBPs (Blanc et al., 2012; Raymond and Morgan-Kiss, 2013, 2017; Mock et al., 2017), mirroring the pattern of ice-associated bacteria and fungi (Margesin et al., 2008). Genome sequencing of the polar diatom *Fragilariopsis cylindrus* identified 11 IBPs (Mock et al., 2017), almost as many as found in ICE-L (12) (Zhang et al., 2020). The IBPs of UWO241 and ICE-L show a surprising degree of similarity with each other as evidenced in the phylogenetic analysis (Figure 3B), especially given that these two algae were isolated from locations that are more than 2500 km apart (Figure 1A). *Chlamydomonas* sp. ICE-MDV, a close relative of ICE-L and a resident of Lake Bonney (Figures 1A, 1C, and 1D), currently holds the record for the greatest number of IBP isoforms (50) in a green alga (Raymond and Morgan-Kiss, 2017). In all these examples, the IBPs are believed to have been acquired from bacteria via HGT, and their existence is thought to be an adaptation to polar environments (Raymond and Kim, 2012). It might seem obvious why a species that lives in the Antarctic would acquire IBPs, which can have ice recrystallization inhibition activities and, thus, protect cells from freezing damage (Davies, 2014). However, the potential benefits bestowed upon UWO241 and ICE-MDV by having these genes is not immediately clear. Unlike ICE-L, UWO241 does not live on ice or snow (Morgan-Kiss et al., 2006) but deep within lake water, which remains at ~5°C year-round (this is also true for ICE-MDV). We do not know the evolutionary history of UWO241 or how long it has been isolated in Lake Bonney, meaning the presence of IBPs could be a remnant of an ancestral lifestyle involving close association with ice and snow.

Genome evolution in a permanently ice-covered Antarctic lake

One must be mindful not to instantly invoke positive selection when trying to explain the evolution of genomic architecture (Lynch, 2007; Brunet and Doolittle, 2018). It is tempting to propose that pervasive gene duplication within the UWO241 genome is an adaptation to life in Lake Bonney. But one could also reason that these features are neutral (or slightly deleterious) outcomes of random genetic events, such as the whims of selfish elements. As with many aspects of molecular evolution, the truth likely falls somewhere in-between these two extremes.

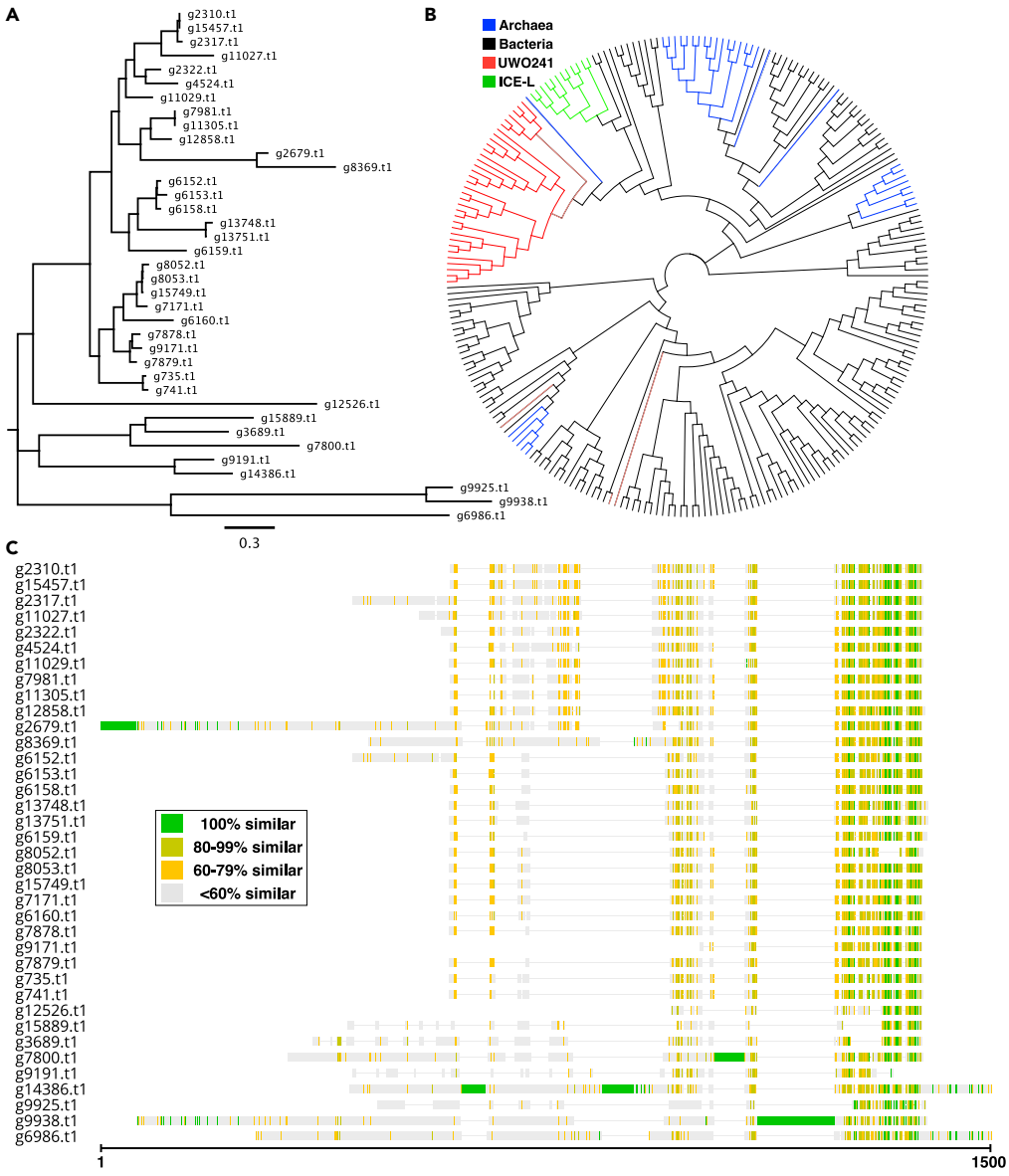


Figure 3. Ice-binding proteins from UWO241

(A) Maximum likelihood (ML) phylogenetic tree based on the amino acid alignments of 37 IBPs in UWO241.

(B) ML phylogenetic relationships of IBPs in UWO241 (red), ICE-L (green), Archaea (blue), and bacteria (black).

(C) Amino acid alignment of 37 IBPs in UWO241 via Clustal Omega, version 1.2.4, using default parameters.

It is our belief that the underlying mechanisms behind the duplications within the UWO241 nucDNA, be it retrotransposition and/or other processes, are neutral or even maladaptive. Likewise, we contend that most of the observed duplicates in the genome, such as those encoding reverse transcriptases, were fixed through random genetic drift, perhaps exacerbated by the hermetic environment of Lake Bonney. (Unfortunately, there are no data on the effective population size of UWO241 and how it compares to that of other green algae, but chlorophytes appear to be relatively rare in Antarctic lake autotrophic communities (Dolhi et al., 2015)). But if enough duplicates are generated, it stands to reason that eventually one will arise resulting in an increase in fitness. For instance, if an increase in dosage of a particular gene is beneficial, then the duplication of this gene could be fixed (or at least maintained after the fact) by positive selection (Innan and Kondrashov, 2010; Kondrashov, 2012). This is arguably the best explanation for the existence of the *petf* duplicates (Cvetkovska et al., 2018), as well as some of the other HSDs in UWO241, including the IBP genes. However, more work is needed, including additional genome sequences from Moewusinia algae, especially close relatives of UWO241, before one can

definitively say if adaptation to an extreme environment is contributing to the retention of HSDs in UWO241. It is noteworthy in this context that neither the UWO241 mitochondrial or chloroplast genomes (Cvetkovska et al., 2019) contain duplicated genes or retroelement-like sequences.

Gene duplication is increasingly being identified as a means of adaptation to extreme environments (Kondrashov, 2012; Qian and Zhang, 2014). Moreover, duplication events resulting in increased gene dosage are known to play a key role in the initial retention of duplicate genes (Innan and Kondrashov, 2010). The data presented here add to this theme. But, again, it is not necessarily true that the infrastructure responsible for generating putatively beneficial duplications is adaptive. Rather, something neutral can sometimes give rise to something useful, which we think is the case for UWO241. The question remains: what precise molecular mechanism(s) are causing genetic duplications in this alga? We favor an RT-based model because of the close association between RTs and duplicates in the genome, as well as the preponderance of reverse transcriptases. But other models are possible. If RTs are contributing to gene duplications in UWO241, then this could help explain the general upheaval we observed throughout the genome but also raises questions about how the HSDs acquired functional regulatory regions—retro-duplication does not typically include regulatory elements but they can be acquire by RTs via other means (Kubiak and Makałowska, 2017). Moreover, RT insertions can also alter the function of nearby genes, which, in turn, can have a cascade of effects (Carelli et al., 2016; Conrad and Antonarakis, 2007).

Remarkably, similar evolutionary processes appear to be operating in the ICE-L genome, in which gene duplication, potentially driven by RTs, has led to large expansions in various gene families, including IBP genes (Zhang et al., 2020), as well as HSDs (265 duplicates covering 717 gene copies) (Figure 1D; Data S4). Many of the HSDs in ICE-L have similar functions to those in UWO241 (Data S4). *C. eustigma*, the closest relative of UWO241 for which a draft genome sequence is available, also has a considerable number of gene duplicates (276), which could be contributing to its survival in an extremely acidic environment (Hirooka et al., 2017). The UWO241, ICE-L, and *C. eustigma* genomes stand in contrast to other explored green algal nucDNAs, which do not have large numbers of HSDs. Indeed, when the same bioinformatics procedures used to identify and classify HSDs in UWO241 were carried out on available chlamydomonadalean genomes, small to moderate numbers of gene duplications were identified (Figures 1D and S4), which is consistent with previous analyses of these genomes. It will be especially interesting to see if ICE-MDV—which like UWO241 lives deep within the water column of Lake Bonney—also harbors expanded gene families and HSDs. Whatever the result, Antarctic lakes have a lot to teach us about genome evolution at the extremes of life.

Limitations of the study

A large number of RTs and rampant gene duplication can cause errors during genome assembly (Zimin et al., 2017). We performed multiple iterations of the UWO241 assembly, using different protocols and algorithms, and are confident that the available draft genome sequence in GenBank is of good quality. The HSDs, in particular, are supported by RNA-seq, meaning there exists a specific transcript corresponding to each duplicate gene. But given the massive extent of duplications in the UWO241 genome, it is likely that some regions were misassembled, especially segments of duplicated noncoding DNA, and will need to be resolved through subsequent sequencing projects. That said, the overall conclusions presented here should not be affected.

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, David R. Smith (dsmit242@uwo.ca).

Materials availability

This study did not generate new reagents or other materials.

Data and code availability

The assembled genome sequences and the raw sequencing data of UWO241 are deposited at the US National Center for Biotechnology Information (NCBI) database under BioProject accession PRJNA547753,

nucleotide accession VFSX00000000, and BioSample accessions SAMN11975472 and SAMN11975511. This study did not generate new code.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102084.

ACKNOWLEDGMENTS

MC, NPAH, and DRS are supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank Bojian Zhong and Jinlai Miao for sharing the ICE-L genome sequences, as well as Rory Craig for useful discussion on TE curation. We also thank Yining Hu for her assistance with computer programming.

AUTHOR CONTRIBUTIONS

The study was conceptualized by MC, NPAH, and DRS. The data were analyzed by MC and XZ. DRS and XZ drafted the manuscript and all authors commented to produce the manuscript for peer review.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 6, 2020
Revised: December 8, 2020
Accepted: January 14, 2021
Published: February 19, 2021

REFERENCES

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., and Lucas, S. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol. *13*, 1–12.

Brunet, T., and Doolittle, W.F. (2018). The generality of constructive neutral evolution. Biol. Philos. *33*, 2.

Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. *26*, 301–314.

Casola, C., and Betrán, E. (2017). The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? Genome Biol. Evol. *9*, 1351–1373.

Catania, F., and Lynch, M. (2008). Where do introns come from? PLoS Biol. *6*, e283.

Conrad, B., and Antonarakis, S.E. (2007). Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu. Rev. Genomics Hum. Genet. *8*, 17–35.

Cvetkovska, M., Orgnero, S., Hüner, N.P., and Smith, D.R. (2019). The enigmatic loss of light-independent chlorophyll biosynthesis from an Antarctic green alga in a light-limited environment. New Phytol. *222*, 651–656.

Cvetkovska, M., Szyszka-Mroz, B., Possmayer, M., Pittock, P., Lajoie, G., Smith, D.R., and Hüner, N.P. (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. New Phytol. *219*, 588–604.

Davies, P.L. (2014). Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. Trends Biochem. Sci. *39*, 548–555.

Demchenko, E., Mikhailyuk, T., Coleman, A.W., and Pröschold, T. (2012). Generic and species concepts in Microglena (previously the *Chlamydomonas monadina* group) revised using an integrative approach. Eur. J. Phycol. *47*, 264–290.

Dolhi, J.M., Teufel, A.G., Kong, W., and Morgan-Kiss, R.M. (2015). Diversity and spatial distribution of autotrophic communities within and between ice-covered Antarctic lakes (McMurdo Dry Valleys). Limnol. Oceanogr. *60*, 977–991.

Dore, J.E., and Priscu, J.C. (2001). Phytoplankton phosphorus deficiency and alkaline phosphatase activity in the McMurdo Dry Valley lakes, Antarctica. Limnol. Oceanogr. *46*, 1331–1346.

Hirooka, S., Hirose, Y., Kanesaki, Y., Higuchi, S., Fujiwara, T., Onuma, R., Era, A., Ohbayashi, R., Uzuka, A., and Nozaki, H. (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proc. Natl. Acad. Sci. U S A *114*, 8304–8313.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Jąkalski, M., Takeshita, K., Deblieck, M., Koyanagi, K.O., Makałowska, I., Watanabe, H., and Makałowski, W. (2016). Comparative genomic analysis of retrogene repertoire in two green algae *Volvox carteri* and *Chlamydomonas reinhardtii*. Biol. Direct *11*, 1–12.

Kalra, I., Wang, X., Cvetkovska, M., Jeong, J., Mchargue, W., Zhang, R., Hüner, N., Yuan, J.S., and Morgan-Kiss, R. (2020). Chlamydomonas sp. UWO 241 exhibits high cyclic electron flow and rewired metabolism under high salinity. Plant Physiol. *183*, 588–601.

Kang, L.-F., Zhu, Z.-L., Zhao, Q., Chen, L.-Y., and Zhang, Z. (2012). Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. BMC Evol. Biol. *12*, 128.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. *9*, 605–618.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B *279*, 5048–5057.

Kubiak, M.R., and Makałowska, I. (2017). Protein-coding genes' retrocopies and their functions. Viruses *9*, 80.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. U S A *104*, 8597–8604.

Margesin, R., Schinner, F., Marx, J.-C., and Gerday, C. (2008). Psychrophiles: From Biodiversity to Biotechnology (Springer Verlag).

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., and Maréchal-Drouard, L. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

Mock, T., Otillar, R.P., Strauss, J., Mcmullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., and Ward, B.J. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature *541*, 536–540.

Morgan-Kiss, R.M., Priscu, J.C., Pocock, T., Gudynaite-Savitch, L., and Huner, N.P. (2006). Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. Microbiol. Mol. Biol. Rev. *70*, 222–252.

Morita, R.Y. (1975). Psychrophilic bacteria. Bacteriol. Rev. *39*, 144–167.

Nakada, T., Misawa, K., and Nozaki, H. (2008). Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Mol. Phylogenet. Evol. *48*, 281–291.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol. *171*, 2294–2316.

Pocock, T., Lachance, M.A., Pröschold, T., Priscu, J.C., Kim, S.S., and Huner, N.P. (2004). Identification of a psychrophilic green alga from Lake Bonney Antarctica: *Chlamydomonas raudensis* ETTL. (UWO241) chlorophyceae. J. Phycol. *40*, 1138–1148.

Possmayer, M., Gupta, R.K., Szyszka-Mroz, B., Maxwell, D.P., Lachance, M.A., Hüner, N.P., and Smith, D.R. (2016). Resolving the phylogenetic relationship between *Chlamydomonas* sp. UWO 241 and *Chlamydomonas raudensis* SAG 49.72 (Chlorophyceae) with nuclear and plastid DNA sequences. J. Phycol. *52*, 305–310.

Qian, W., and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. Genome Res. *24*, 1356–1362.

Raymond, J.A., and Kim, H.J. (2012). Possible role of horizontal gene transfer in the colonization of sea ice by algae. PLoS One *7*, e35968.

Raymond, J.A., and Morgan-Kiss, R. (2013). Separate origins of ice-binding proteins in Antarctic *Chlamydomonas* species. PLoS One *8*, e59186.

Raymond, J.A., and Morgan-Kiss, R. (2017). Multiple ice-binding proteins of probable prokaryotic origin in an Antarctic lake alga, *Chlamydomonas* sp. ICE-MDV (Chlorophyceae). J. Phycol. *53*, 848–854.

Szcześniak, M.W., Ciomborowska, J., Nowak, W., Rogozin, I.B., and Makałowska, I. (2011). Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol. Biol. Evol. *28*, 33–37.

Wikimedia-Commons. (2020). Wikimedia commons, the free media repository. https://commons.wikimedia.org/wiki/Main_Page.

Zhang, C., Gschwend, A.R., Ouyang, Y., and Long, M. (2014). Evolution of gene structural complexity: an alternative-splicing-based model accounts for intron-containing retrogenes. Plant Physiol. *165*, 412–423.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298.

Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X., and Wang, W. (2020). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Curr. Biol. *30*, 1–12.

Zhu, Z., Zhang, Y., and Long, M. (2009). Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. Plant Physiol. *151*, 1943–1951.

Zimin, A.V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J., and Salzberg, S.L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. *27*, 787–792.

iScience, Volume *24*

Supplemental Information

Draft genome sequence of the Antarctic green alga *Chlamydomonas* sp. UWO241

Xi Zhang, Marina Cvetkovska, Rachael Morgan-Kiss, Norman P.A. Hüner, and David Roy Smith

Supplemental Figures

A

Cells measured by Flow cytometry	Nuclear DNA content (1 pg = 978 Mb)	Estimated genome size (Mb)	Standard error of mean genome size (Mb)
Mixed cell culture of UWO241	0.53 pg/2C	259	2

B (a)

k-mer length	65	70	75
Total *k*-mer count	4574846901	3783302136	3039573138
Coverage	23	18	14
Genome size (Mb)	198.9	210.2	217.1

C

Genome statistics	***Chlamydomonas*. sp UWO241**
Genome size (Mb)	211.6
Scaffold N50 (Mb)/L50	0.37/165
GC (%)	60.61
Number of protein coding genes	16325
Gene density (genes/Mb)	77.2
mean gene length	8058
Average exon length (bp)	1563
% of genome covered by genes	62.8
% of genome covered by CDS	12.2
Average intron per gene	10.11
Average intron length (bp)	934
Genome completness by BUSCO	85%

B (b)

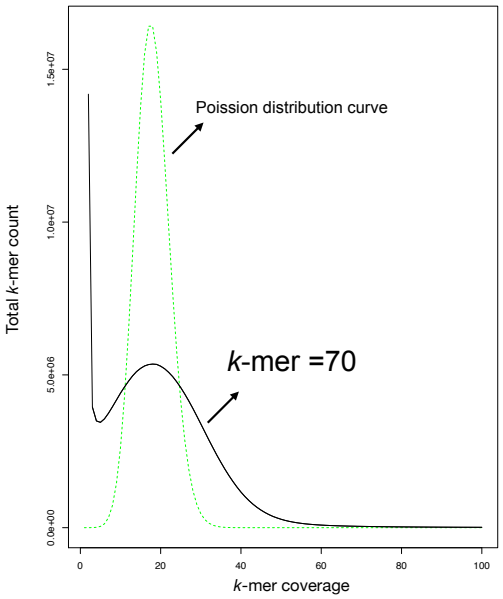


Figure S1. UWO241 nuclear genome size and content. Related to Figure 1. (A) Estimated genome size based on flow cytometry. (B) (a) Estimated genome size based on *k*-mer lengths; (b) *k*-mer histogram of UWO241 (X-axis is the number of times a given *k*-mer was observed in the UWO241 sequencing data and Y-axis is the total number of *k*-mers with a given *k*-mer coverage). (C) Nuclear genome content.

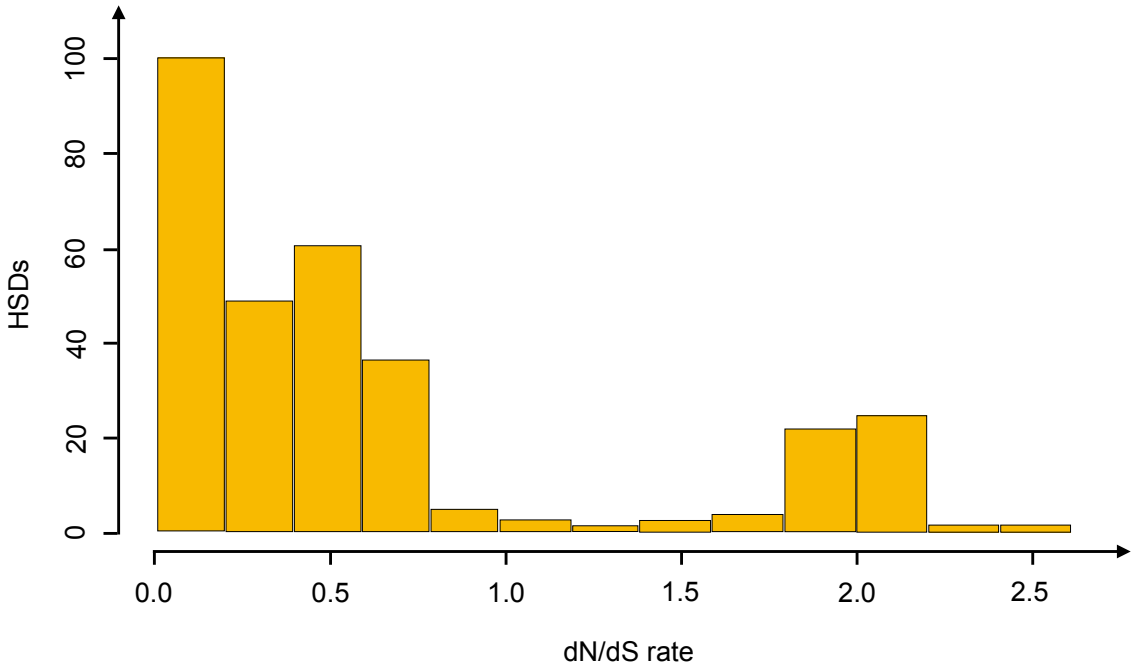
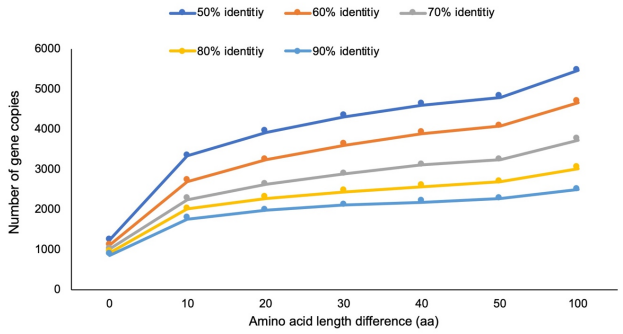


Figure S2. The distribution of nonsynonymous to synonymous substitution rates (dN/dS) among 316 HSDs in UWO241. Related to Table 2.

A



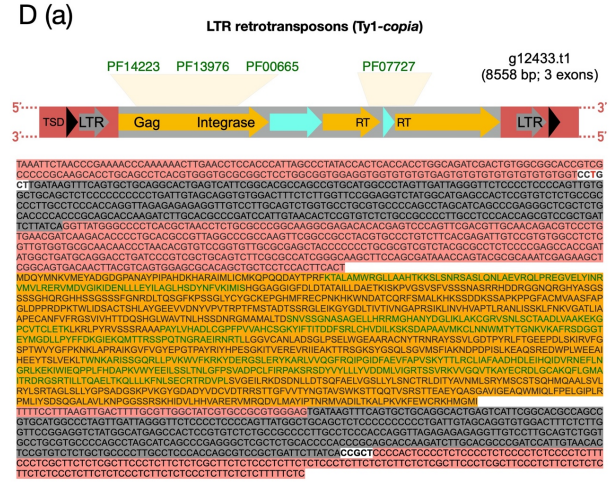
B

	50% identitiy	60% identitiy	70% identitiy	80% identitiy	90% identitiy
100 aa	5453	4654	3719	3012	2476
50 aa	4783	4061	3228	2672	2257
40 aa	4594	3879	3088	2565	2181
30 aa	4310	3596	2875	2438	2091
20 aa	3912	3224	2614	2270	1973
10 aa	3320	2699	2234	2012	1339
0 aa	1223	1097	999	926	859

C

K-mer length	65	70	75
Genome size (Mb)	198.9	210.2	217.1
Single copy region (Mb)	165.6	182.7	186.5
Duplicate copy region (Mb)	33.3	27.5	30.6
HSD duplicates size (90%) (Mb)	4.4	4.4	4.4
Partial duplicate zone size (Mb)	28.9	23.1	26.2
Partial duplicate zone porptotation	0.14	0.11	0.12

D (a)



D (b)

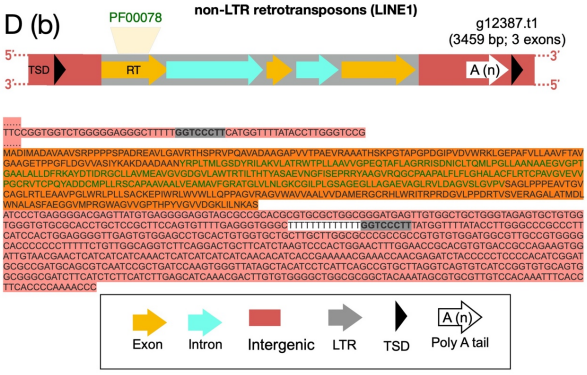


Figure S3. Partial gene duplicates, retrogenes, and retrotransposons in UWO241. Related to Figure 2. (A) Line graph of putative duplicate genes based on various pairwise amino acid identity thresholds and amino acid length differences; the X-axis indicates the amino acid length differences among duplicates and the Y-axis is the number of gene copies. (B) Number of duplicate gene copies at different amino acid pairwise identities and length thresholds. (C) Approximation of the proportion of duplicated regions in the UWO241 genome. (D) (a) Example of and LTR-retrotransposon (Ty1-*copia*) in the UWO241 genome; the retrotransposon is flanked by long terminal repeats (LTRs, grey) and short black triangles indicating target site duplications (TSDs, black). (D) (b) Example of a non-LTR retrotransposon (LINE1), which is terminated by a 3'-poly(A) tail (A(n), white arrow). The Pfam domains (green) are PF14223: gag-polypeptide of LTR "*copia*-type"; PF13976: gag-pre-integrase domain; PF00665: integrase core domain; PF07727: reverse transcriptase; PF00078: reverse transcriptase.

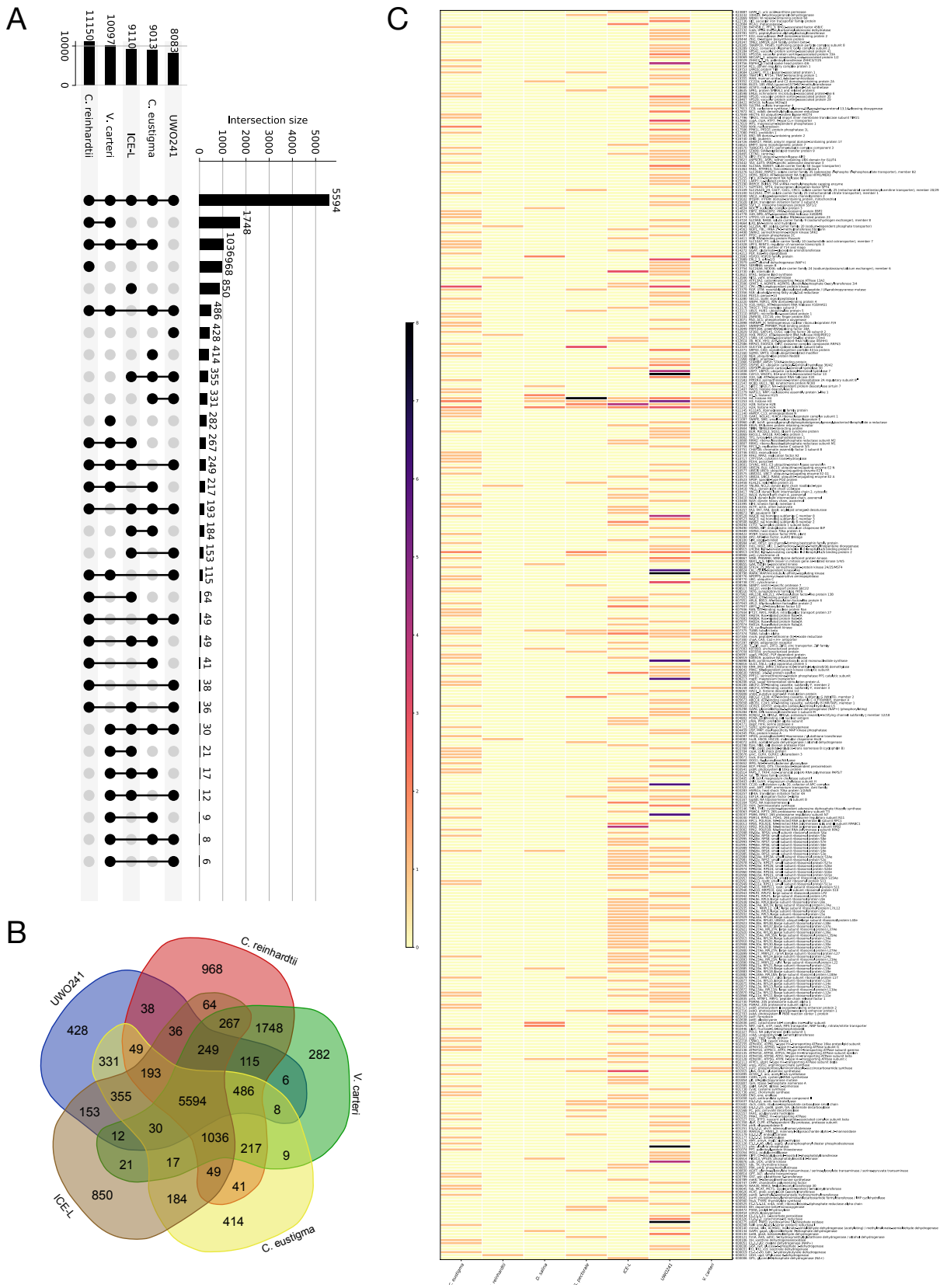

A
11150
10097
9110
9013
8083
C. reinhardtii
V. carteri
ICE-L
C. eustigma
UWO241
Intersection size
0
1000
2000
3000
4000
5000
10000
5594
1748
1036
968
850
486
428
414
355
331
282
267
249
217
193
184
153
115
64
49
49
41
38
36
30
21
17
12
9
8
6
B
C. reinhardtii
V. carteri
ICE-L
C. eustigma
UWO241
968
64
267
1748
38
36
428
49
249
115
282
331
193
6
355
5594
486
153
8
12
30
1036
217
9
21
17
49
41
850
184
414
C
C. eustigma
reinhardtii
D. salina
T. pseudonana
ICE-L
UWO241
V. carteri

Figure S4. Comparative genomic analyses across selected chlamydomonadalean algae. Related to Figure 2 and Table 2. (A) and (B) UpSet plot and Venn diagram, respectively, displaying unique and shared gene families among UWO241 and its chlamydomonadalean relatives. (C) Number of HSDs across various chlamydomonadalean species grouped based on their KEGG functional category.

TRANSPARENT METHODS

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strain and culture conditions

UWO241 is available from the National Center for Marine Algae and Microbiota (NCMA) under strain number CCMP 1619. The strain used for genome sequencing is the original isolate, obtained directly from Priscu (Pocock et al., 2004). UWO241 was cultured at 5 °C and aerated continuously with ambient air filtered by a 0.2 μm filter in 250 ml glass growth tubes suspended in thermo-regulated aquaria. All cultures were grown at a continuous light of 100 μmol photons m^{-2} s^{-1}, measured with a quantum sensor attached to a radiometer (Model LI-189; Li-Cor, Lincoln, NE, USA). Cultures were grown to mid-log phase prior to harvesting.

METHOD DETAILS

DNA and RNA extraction and library construction

Genomic DNA (gDNA) for Illumina HiSeq2000 sequencing was extracted using the Qiagen Plant DNeasy Maxi Kit (Qiagen), following manufacturer's instructions. UWO241 was harvested by centrifugation (6000 g, 5 min, 4 °C), flash-frozen in liquid nitrogen, and stored at 80 °C. After extraction, the gDNA was purified by ethanol precipitation using standard methods and resuspended in 10 mM Tris, pH 7.5. DNA quality was monitored using wavelength absorbance scan and electrophoresis on a 1% (w/v) TBE agarose gel.

For single-molecule, real-time (SMRT) sequencing (Pacific Biosciences, Menlo Park, CA, USA), gDNA was extracted using a modified CTAB protocol. In short, cell pellets were resuspended in 1 ml of lysis buffer (50 mM Tris-HCl at pH 8.0, 200 mM NaCl, 20 mM EDTA, 2% w/v SDS, and 20 mg/ml Proteinase K) and mixed by inversion. Equal volume of pre-heated CTAB buffer (2% w/v) CTAB, 1.4 M sodium chloride, 20 mM EDTA, 100 mM Tris, 2% (w/v) polyvinylpyrrolidone (M.W. 40000), 1% (v/v) β-mercaptoethanol, pH 8.0) was added, and the cells were incubated at 65 °C for 30 min, followed by centrifugation (14,000g, 5 min) to remove the insoluble materials. The supernatant was treated with RNase A (100 mg/ml) for 30 min at 37 °C, and nucleic acids were extracted 2x with equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). The extract was centrifuged (16,000g, 10 min) and nucleic acids were precipitated with 1x volume of ice-cold isopropanol and incubated at -20 °C for 1 hour. The samples were centrifuged (16,000g, 15 min, 4 °C) and the pellet was washed 3x with ice-cold 70% (v/v) ethanol. DNA was precipitated with 1/10th volume 3M sodium acetate (pH 5.2) and 2 volumes 100% ethanol, and then samples were incubated at -20 °C for 1hour, centrifuged (16000g, 4 °C, 30 min), and the resulting DNA pellets washed with 70% (v/v) ethanol. The pellets were air-dried and resuspended in 10 mM Tris (pH 7.8) by incubating them for 24 hours at 4 °C.

RNA sequencing (RNA-seq) of UWO241 was performed using 125 nt paired-end (PE) reads on an Illumina HiSeq 2500 v4 sequencing platform. Three biological replicate cultures of UWO241

were grown at 15 °C. Algal cells were harvested by centrifugation (6,000g, 5 min, 4 °C), flash frozen in liquid nitrogen, and stored at -80 °C. RNA was isolated using a modified CTAB protocol (Possmayer et al., 2016) and sequenced at the Génome Québec Innovation Centre (Montreal, QC, Canada). Total RNA was quantified using a NanoDrop Spectrophotometer ND-1000 (NanoDrop Technologies, Inc.) and its integrity was assessed using a 2100 Bioanalyzer (Agilent Technologies). Libraries were generated from 250 ng of total RNA using the TruSeq stranded mRNA Sample Preparation Kit (Illumina), as per manufacturer's recommendations. Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument.

Library construction and sequencing

Genomic HiSeq 2000 sequencing was performed at the Princess Margaret Genomics Centre (Toronto, ON, Canada), using 101-cycle PE reads at 100x coverage. DNA was fragmented using a Covaris M220 Focused-Ultrasonicator (Covaris Inc., Woburn, MA, USA) and libraries were constructed with the TruSeq DNA HT Sample Preparation Kit (FC-121-2003; Illumina, San Diego, CA, USA). PacBio SMRT sequencing was performed by Génome Québec on an RSII instrument, using 19 cells at 81x coverage. 7.5 µg of high-molecular-weight gDNA was sheared using the Covaris g-TUBES (Covaris Inc.). DNA libraries were prepared using the SMRTbell Template Prep Kit 1.0 reagents (Pacific Biosciences). The DNA library was size-selected on a BluePippin system (Sage Science Inc., Beverly, MA, USA) using a cut-off range of 10-50 kb.

Estimation of genome size

The nuclear genome size of UWO241 was estimated using *k*-mer analysis and flow cytometry. Approximately ~30 Gb of high-quality, short-insert reads (250 bp) were used to estimate genome size via the *k*-mer analysis tool Jellyfish (Marçais and Kingsford, 2011). The *k*-mer frequency followed a Poisson distribution. The *k*-mer depth (i.e., mean coverage) was divided by the total *k*-mer number, giving a genome-size estimate of 210Mbp (± 10Mbp; mean ± standard error) when using a default *k*-mer size 65, 70 and 75. The genome size estimation via *k*-mer is followed through the tutorial with the link (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/).

Flow cytometry predicted the UWO241 genome size to be 250 Mbp (± 2Mbp; mean ± standard error), following a modified protocol by Arumuganathan and Earle (Arumuganathan and Earle, 1991b). Briefly, intact nuclei were suspended in $MgSO_4$ buffer mixed with DNA standards and stained with propidium iodide (PI) in a solution containing DNAase-free RNAase (Arumuganathan and Earle, 1991a). Fluorescence intensities of the stained nuclei were measured by a flow cytometer. Values for nuclear DNA content were estimated by comparing fluorescence intensities of the nuclei of UWO241 with those of various internal DNA standards, including nuclei from Chicken Red blood cells (2.5 pg/2C), *Glycine max* (2.45 pg/2C), *Oryza sativa* cv Nipponbare (0.96 pg/2C), and *Arabidopsis thaliana* (0.36 pg/2C). Specifically, for flow cytometric

analysis, one mL of UWO241 culture was placed in microfuge tubes and centrifuged for 5 sec. The pellet was suspended by vortexing vigorously in 0.5 mL solution containing 10 mM $MgSO_4.7H_2O$, 50mM KCl, 5 mM Hepes, pH 8.0, 3 mM dithiothreitol, 0.1 mg / mL propidium iodide, 1.5 mg / mL DNAse free RNAse (Rhoche, Indionapolis, IN) and 0.25% Triton X-100. The suspended nuclei were withdrawn using a pipettor, filtered through 30-µm nylon mesh, and incubated at 37 °C for 30 min before flow-cytometric analysis. Suspensions of sample nuclei was spiked with suspension of standard nuclei (prepared in above solution) and analyzed with a FACScalibur flow cytometer (Becton-Dickinson, San Jose, CA). For each measurement, the propidium iodide fluorescence area signals (FL2-A) from 1000 nuclei were collected and analyzed by CellQuest software (Becton-Dickinson, San Jose, CA) on a Macintosh computer (Dickinson and Dickinson, 1998). The mean position of the G0/G1 (nuclei) peak of the sample and the internal standard were determined by CellQuest software. The mean nuclear DNA content of each plant sample, measured in picograms, was based on 1000 scanned nuclei.

Genome assembly

The nuclear genome of UWO241 was assembled *de novo* using Illumina and PacBio SMRT sequencing reads. The Illumina read quality was evaluated using FastQC v0.11.8 (Andrews, 2010), and the PacBio sequencing reads were assessed via the error-correction step of Canu v1.7.1 (Koren et al., 2017). The hybrid *de novo* assembly was carried out with MaSuRCA v3.3.2 (Zimin et al., 2017), using an automatically determined *k*-mer size (i.e., GRAPH_KMER_SIZE = auto), which computes the optimal size based on the read data and GC content; a cgwErrorRate of 0.15; and a KMER_COUNT_THRESHOLD of 1. Scaffolding and gap-filling algorithms were then applied to all hybrid-assembled contigs to extend the length of the assembly and to minimize mismatches. SSPACE v3.0 (Boetzer et al., 2010) was used to extend and scaffold pre-assembled contigs by using Illumina PE libraries. GapFiller v2.1.1 (Boetzer and Pirovano, 2012) was used to close the gaps ('N') in the scaffolds by mapping with long PacBio reads. The genome assembly was further polished with highly accurate Illumina reads via Pilon v1.22 (Walker et al., 2014). Assemblies of the plastid and mitochondrial genomes were produced independently (Cvetkovska et al., 2019). The Illumina HiSeq transcriptomic data were *de novo* assembled via Trinity v2.8.4 (Haas et al., 2013). Adapters and low-quality bases were trimmed from each RNA-seq dataset using Trimmomatic v0.38 (Bolger et al., 2014). Genome assembly metrics were generated using QUAST v5.0.0 (Gurevich et al., 2013).

De novo repeat finding and repeat masking

A *de novo* repeat library was created with RepeatModeler v1.0.8 (Smit and Hubley, 2008), RepeatScout v1.0.5 (Price et al., 2005), LTR_FINDER (Xu and Wang, 2007), and LTR_retriever (Ou and Jiang, 2018) using default parameters. Unknown elements were screened with BLASTX (Altschul et al., 1997) (E-value < 1e-5) against the UniRef90 database (Suzek et al., 2015) (subset Viridiplantae) and removed from the repeat library if necessary. The repeat library of UWO241 was used by RepeatMasker (4.0.7) (rmblastn version 2.2.27+) (Tarailo-Graovac and Chen, 2009)

to mask the repetitive elements in the assembly. The masked regions were further inspected for overlaps with UWO241 RNA-seq transcripts via GENEIOUS v.10.1 (Biomatters Ltd, Auckland, New Zealand) (Kearse et al., 2012). Considering some genes, such as TE-related ones, can partially overlap with repeat regions, it is not uncommon to have some "noise" when inspecting the masked regions. RepeatMasker (Tarailo-Graovac and Chen, 2009) allows for a soft-masked genome option to help prevent from over-masking.

Gene prediction

Coding regions were annotated by incorporating RNA-seq data with the *ab initio* gene prediction tool AUGUSTUS v3.0.3 (Stanke et al., 2008). RNA-seq transcripts were fed into the pipeline of AUGUSTUS as hints using the "--UTR=on" and "--alternatives-from-evidence=true" options. UTR flag was set to "on" to perform untranslated region annotations. The alternative-evidence flag was set to "true" to predict alternative splicing. The training sets of AUGUSTUS were acquired from the first run of EVidenceModeler (EVM) (Haas et al., 2008) gene models. The extrinsic evidence for EVM was acquired from transcript alignments and homolog-based predictions. The RNA-seq data were first used to reconstruct the transcripts via Trinity v2.8.4 (Haas et al., 2013), then the transcripts alignments for EVM were created using PASA v2.3.3 (Haas et al., 2003). To create the evidence of homolog-based predictions, the protein sequences of five closely related species (*Chlamydomonas reinhardtii* (Merchant et al., 2007), *Gonium pectorale* (Hanschen et al., 2016)*, Chlamydomonas eustigma* (Hirooka et al., 2017)*, Dunaliella salina* (Polle et al., 2017) and *Volvox carteri* (Prochnik et al., 2010)) were downloaded from JGI (https://phytozome.jgi.doe.gov/pz/portal.html) or NCBI (https://www.ncbi.nlm.nih.gov) database. The evidence of protein alignments for EVM were created with Exonerate (Slater and Birney, 2005), seeded by Diamond (Buchfink et al., 2015). The list of numeric weight values was set to default for each type of "evidence" for EVM.

Functional annotations of protein-coding genes were obtained from the best BLAST hit by BLASTP (E-value < 1e-5) against SwissProt (Consortium, 2019), TrEMBL (Boeckmann et al., 2003), and NCBI NR databases (non-redundant protein sequence database with entries from GenPept, SwissProt, PIR, PDF, PDB, and RefSeq). We developed a tool called NoBadWordsCombiner v1.0 (Zhang et al., 2020c), which can automatically merge the BLAST results from the databases of SwissProt (Consortium, 2019), TrEMBL (Boeckmann et al., 2003) and NCBI NR databases. More importantly, it can strengthen the gene definition by filtering those protein function descriptions containing 'bad words', such as hypothetical and uncharacterized proteins. GENEIOUS v.10.1 (Biomatters Ltd, Auckland, New Zealand) was used to visualize the gene models and manually trim short gene models. The gene models were manually filtered if genes contained internal stop codons, deduced protein sequences less than 35 amino acids, or coding regions with > 70% of elements from low complexity regions and simple repeats. Pfam domains were annotated using InterProScan (v4.7) (Zdobnov and Apweiler, 2001), which integrates predictive information about protein function from a number of partner resources, such

as the InterPro (Quevillon et al., 2005) and Pfam (Finn et al., 2014) databases. Gene Ontology (GO) terms (Ashburner et al., 2000) for each gene were retrieved from the corresponding InterPro or Pfam descriptions. Gene sets were mapped to KEGG (Kanehisa and Goto, 2000) pathways to identify the best match classification for each gene. Genome annotation quality was evaluated by BUSCO (Simão et al., 2015), which gave a quantitative measures for single-copy orthologous genes from the dataset Chlorophyta odb10 (Zdobnov et al., 2017). The tRNA genes were predicted by tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997) using default parameters for eukaryotes.

Comparative genomic analyses

Protein sequences from the nuclear genomes of 7 green algae belonging to the Chlorophyta (*C. reinhardtii* (Merchant et al., 2007), *G. pectorale* (Hanschen et al., 2016), *C. eustigma* (Hirooka et al., 2017), *D. salina* (Polle et al., 2017), *V. carteri* (Prochnik et al., 2010), *Chlamydomonas* sp. ICE-L (Zhang et al., 2020d) were used to construct homologous gene clusters (orthogroups) by OrthoFinder v2.1.2 (Emms and Kelly, 2015). The longest transcript of each gene was retained to remove redundancy resulting from alternative splicing variations, and genes encoding protein sequences shorter than 50 amino acids were filtered to exclude putative fragmented genes. Orthogroups with single-copy genes shared by all 7 genomes were retained for further analyses. The single-copy genes were retrieved to create a phylogenetic tree. The Venn diagram is created via the online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). Maximum likelihood trees were generated using RAxML v7.0.4 (Stamatakis et al., 2004) with the PROTCATJTT model.

Highly similar duplicate genes (HSDs) predictions

A protein BLAST (Altschul et al., 1997) of the UWO241 gene models against themselves (E-value cut-off 10^{-5}) was filtered to only those with near-identical protein lengths (within 10 amino acids) and ≥90% pairwise identities (Data S2b). This gave a list of highly similar duplicates (HSDs). The deduced amino acid sequences of the HSDs were searched against the KEGG (Kanehisa and Goto, 2000) and Pfam databases (Finn et al., 2014), providing a functional breakdown. To extensively identify HSDs with high accuracy and reliability, we developed a web-based tool HSDFinder (http://hsdfinder.com) (Zhang et al., 2020b), which we also used to predict HSDs in other chlorophyte algae. The predicted results are documented in the database of HSDatabase (http://hsdfinder.com/database/) (Zhang et al., 2020a). Using HSDFinder, users have the option to employ different parameters (from 30% to 100% identity and from within 0-100 aa variances) for identifying HSDs.

Substitution rate analysis of highly similar duplicate genes (HSDs)

The protein sequences of each HSD gene copy were aligned using Clustal Omega v1.2.4 (Sievers et al., 2011), poorly aligned regions were trimmed with trimAl v1.4 (Capella-Gutiérrez et al., 2009). Nonsynonymous (dN) and synonymous (dS) substitution rates were calculated for each HSD group by reverse-translating the amino acid alignments to the corresponding codon-based nucleotide alignments using PAL2NAL (Suyama et al., 2006). Both amino acid and codon

sequence alignments were separately concatenated using GENEIOUS v.10.1 (Kearse et al., 2012). Maximum likelihood (ML) phylogenetic trees were inferred based on protein and codon alignments using FastTree v2.1 (Price et al., 2010) with default parameters. We then applied the one-ratio model in the codeml program of PAML v4.9 (Yang, 2007) to estimate the dN/dS substitution rates (ω value) with the parameters "runmode = 0" and "model =0".

Horizontal gene transfer (Ice binding proteins)

Preliminary BLAST analyses (BLASTP, E-value < 1e-5) showed that a small proportion of genes in the UWO241genome had a top hit to sequences from non-green algae sources, suggesting that these genes might have been acquired through horizontal gene transfer (HGT). Several steps were taken to estimate the overall reliability of HGT. We checked all annotated genes based on their non-redundant annotations, and extracted genes with non-plant annotations (i.e., those matching to fungi, bacteria, archaea and virus) as candidate for further analyses. The BLAST protein databases labeled as fungi, bacteria, archaea, and viruses were downloaded from UniProt (https://www.uniprot.org/downloads) and used to perform BLASTP searches with an E-value < 1e-5. The bit-score of the top ten BLAST hits were extracted as the candidate HGT genes for further analysis. The Clustal Omega v1.2.4 (Sievers et al., 2011) was used to align the candidate HGT genes. Each alignment was trimmed to exclude regions where only one of the sequences was present, and maximum likelihood phylogenetic trees were built using FastTree v2.1 (Price et al., 2010) from amino-acids sequences. Gene trees supporting a sister grouping between UWO241 and a non-plant with support value ≥ 80 were retained as putative HGT events. The IBPs sequences from ICE-L were collected from the supplemental data of the ICE-L genome paper (Zhang et al., 2020d). The archaea and bacteria used for phylogenetic analyses with IBPs were also retrieved from ICE-L genome paper.

QUANTIFICATION AND STATISTICAL ANALYSIS

The details of the statistics applied (e.g., genome sequencing, evaluation of assembly accuracy and genome-quality, and gene transcriptome quantification) can be found in the relevant sections of the Methods Details. The phylogenetic trees were reconstructed using FastTree v2.1 (Price et al., 2010) and RAxML v7.0.4 (Stamatakis et al., 2004). Statistical tests were carried out in R (https://www.r-project.org/).

ADDITIONAL RESOURCES

To extensively identify HSDs with high accuracy and reliability, we developed a web-based tool called HSDFinder (http://hsdfinder.com) (Zhang et al., 2020b) to predict HSDs in the eukaryotic genomes.

Supplemental References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. Retrieved from https://www.bioinformatics.babraham.ac.uk.

Arumuganathan, K. & Earle, E. (1991a). Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol. Biol. Rep., 9, 229-241.

Arumuganathan, K. & Earle, E. (1991b). Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep., 9, 208-218.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. & Eppig, J. T. (2000). Gene ontology: tool for the unification of biology. Nat. Genet., 25, 25-29.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'donovan, C. & Phan, I. (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res., 31, 365-370.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. (2010). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27, 578-579.

Boetzer, M. & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol., 13, R56.

Bolger, A. M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114-2120.

Buchfink, B., Xie, C. & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods, 12, 59.

Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 1972-1973.

Consortium, U. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res., 47, D506-D515.

Cvetkovska, M., Orgnero, S., Hüner, N. P. & Smith, D. R. (2019). The enigmatic loss of light-independent chlorophyll biosynthesis from an Antarctic green alga in a light-limited environment. New Phytol., 222, 651-656.

Dickinson, N. B. & Dickinson, B. (1998). CellQuest Software Reference Manual. Becton Dickinson Immunocytometry Systems, San Jose, 1-227.

Emms, D. M. & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol., 16, 157.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L. & Mistry, J. (2014). Pfam: the protein families database. Nucleic Acids Res., 42, 222-230.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072-1075.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B. & Town, C. D. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res., 31, 5654-5666.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B. & Lieber, M. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc., 8, 1494.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R. & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol., 9, R7.

Hanschen, E. R., Marriage, T. N., Ferris, P. J., Hamaji, T., Toyoda, A., Fujiyama, A., Neme, R., Noguchi, H., Minakuchi, Y. & Suzuki, M. (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nat. Commun., 7, 1-10.

Hirooka, S., Hirose, Y., Kanesaki, Y., Higuchi, S., Fujiwara, T., Onuma, R., Era, A., Ohbayashi, R., Uzuka, A. & Nozaki, H. (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proc. Natl. Acad. Sci. U.S.A., 114, 8304-8313.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res., 28, 27-30.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S. & Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28, 1647-1649.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res., 27, 722-736.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res., 25, 955-964.

Marçais, G. & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics, 27, 764-770.

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K. & Maréchal-Drouard, L. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science, 318, 245-250.

Ou, S. & Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol., 176, 1410-1422.

Pocock, T., Lachance, M. A., Pröschold, T., Priscu, J. C., Kim, S. S. & Huner, N. P. (2004). Identification of a psychrophilic green alga from Lake Bonney Antarctica: *Chlamydomonas raudensis* ETTL. (UWO241) Chlorophyceae. J. Phycol., 40, 1138-1148.

Polle, J. E., Barry, K., Cushman, J., Schmutz, J., Tran, D., Hathwaik, L. T., Yim, W. C., Jenkins, J., Mckie-Krisberg, Z. & Prochnik, S. (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announc., 5, 01105-17.

Possmayer, M., Gupta, R. K., Szyszka-Mroz, B., Maxwell, D. P., Lachance, M. A., Hüner, N. P. & Smith, D. R. (2016). Resolving the phylogenetic relationship between *Chlamydomonas* sp. UWO 241 and *Chlamydomonas raudensis* SAG 49.72 (Chlorophyceae) with nuclear and plastid DNA sequences. J. Phycol., 52, 305-310.

Price, A. L., Jones, N. C. & Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics, 21, 351-358.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One, 5, e9490.

Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., Ferris, P., Kuo, A., Mitros, T. & Fritz-Laylin, L. K. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science, 329, 223-226.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res., 33, 116-120.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., Mcwilliam, H., Remmert, M. & Söding, J. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol., 7, 539.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31, 3210-3212.

Slater, G. S. C. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinform., 6, 1-11.

Smit, A. & Hubley, R. (2008). RepeatModeler Open-1.0. Retrieved from http://www.repeatmasker.org.

Stamatakis, A., Ludwig, T. & Meier, H. (2004). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics, 21, 456-463.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics, 24, 637-644.

Suyama, M., Torrents, D. & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res., 34, 609-612.

Suzek, B. E., Wang, Y., Huang, H., Mcgarvey, P. B., Wu, C. H. & Consortium, U. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31, 926-932.

Tarailo-Graovac, M. & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics, 4.10. 1-4.10. 14.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J. & Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One, 9, e112963.

Xu, Z. & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res., 35, 265-268.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol., 24, 1586-1591.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17, 847-848.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., Seppey, M., Loetscher, A. & Kriventseva, E. V. (2017). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res., 45, D744-D749.

Zhang, X., Hu, Y. & Smith, D. R. (2020a). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. Retrieved from http://hsdfinder.com/database/.

Zhang, X., Hu, Y. & Smith, D. R. (2020b). HSDFinder- an integrated tool for predicting highly similar duplicates in eukaryotic genomes. Retrieved from https://github.com/zx0223winner/HSDFinder.

Zhang, X., Hu, Y. & Smith, D. R. (2020c). NoBadWordsCombiner-a tool to integrate the gene function information together without 'bad words' from Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam databases. . Retrieved from https://github.com/zx0223winner/HSDFinder/blob/master/NoBadWordsCombiner.py.

Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X. & Wang, W. (2020d). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Curr. Bio., 30**,** 1-12.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J. & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res., 27**,** 787-792.