The Chlamydomonas nuclear genome

Rory J Craig, Olivier Vallon

Chapter 5: The Chlamydomonas nuclear Genome

Rory J. Craig[1,2] and Olivier Vallon[3]

[1] California Institute for Quantitative Biosciences, UC Berkeley, Berkeley, USA
[2] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
[3] CNRS, Sorbonne Université, UMR7141, Institut de Biologie Physico-Chimique, Laboratory of Chloroplast Biology and Light-Sensing in Microalgae, 75005 Paris, France.

CHAPTER CONTENTS

ABSTRACT

Chlamydomonas has become a premier model organism thanks to the ease with which gene function can be analyzed experimentally. With the advent of genome and transcriptome sequencing, generating an accurate assembly of the genome and correctly annotating the structure of genes has become an essential goal for the *Chlamydomonas* research community. Here, we review the history of the field and present in detail the most recent description of the genome and its genes, to be released during the summer of 2022. We further describe the genome from a structural perspective and discuss genomic variation among *Chlamydomonas* laboratory strains and field isolates. Genomics is an ever-evolving field, and we expect that the description of the genome will be continuously enriched as new tools develop and more strains are analyzed.

I. INTRODUCTION

The *Chlamydomonas* nuclear genome is approximately 111 Mb in length, GC-rich (~64% genome-wide) and arranged on 17 chromosomes ranging from 3.7 Mb to 9.8 Mb. The first draft assemblies were produced in the early to mid 2000s, among the first wave of Sanger-sequenced eukaryotic genomes. However, early versions of the *Chlamydomonas* reference assembly contained many gaps and some misassemblies, unlike many of its near-complete contemporaries e.g. the *Arabidopsis thaliana* and *Oryza sativa* assemblies (Arabidopsis Genome Initiative 2000; Goff et al. 2002). Recent developments have now seen the production of a highly contiguous reference assembly. Similarly, the quality of the gene structural annotations has improved considerably over the last two decades, following advances in sequencing technologies and underlying improvements to the assembly.

This chapter reviews the structure of the genome and of the genes, while Chapter 4 of this volume describes the functions of the encoded proteins. To help the reader understand the context in which previous research was performed and to foresee future developments, we first present the history of the reference assembly and gene model annotations. Major features of genome architecture and gene organization are subsequently described. Finally, we discuss genome evolution and introduce ongoing efforts to produce genomic resources for multiple *Chlamydomonas* strains.

II. THE *CHLAMYDOMONAS* REFERENCE GENOME

A. Versions 1-5: improvements of a draft genome

Chlamydomonas genomics is an ongoing community effort, borne from a shared desire to offer a vibrant research community both a description of the species "blueprint" and a resource to characterize and manipulate individual genes. It was made possible by the extensive genetic analysis data accumulated since the 1950s and the pioneering efforts of a few laboratories who built libraries of genomic DNA (Shrager et al. 2003; Kathir et al. 2003) or cDNA (Section III). In the early 2000s, the DOE Joint Genome Institute (JGI) sequenced, assembled and annotated the genome of the cell wall-less strain CC-503 (*cw92*), chosen at

that time because preparation of high-quality DNA was facilitated by the absence of a cell wall. CC-503 was derived from the *mt*+ strain CC-125 (137c+) by mutagenesis with the methylating agent *N*-methyl-*N*'-nitro-*N*-nitrosoguanidine (Hyams and Davies 1972). The *Chlamydomonas* Genome Project initially used Sanger sequencing of the insert ends of plasmid (2-3 kb or 6-8 kb) and fosmid (35-40 kb) libraries to achieve ~13x coverage of the genome, and subsequently assembled contiguous stretches of sequence (i.e. "contigs") from the overlapping sequence of multiple reads. Contigs were then arranged and linked together with gaps of unknown sequence to form "scaffolds" based on the information obtained from sequencing both ends of inserts. Preceded by two preliminary versions (Grossman et al. 2003), the first high-quality draft assembly (v3) was presented in the landmark genome paper of Merchant et al. (2007). This assembly also utilized additional long-range association between scaffolds made possible by the end-sequencing of two BAC libraries with median insert size 48 kb and 185 kb (University of Minnesota and Exelixis, respectively).

The v3 assembly consisted of 1,557 scaffolds, spanning 120.2 Mb and with 12.5% gaps. The scaffold N50 was 1.7 Mb, meaning that one-half of the cumulative sequence in the assembly was present on scaffolds of this size or longer. However, the contig N50 was only 44.5 kb, indicating an abundance of gaps within the scaffolds. Utilizing decades of molecular and genetic mapping data, the v3 scaffolds were aligned to the 17 *Chlamydomonas* linkage groups (Kathir et al. 2003; Rymarquis et al. 2005), although only 78% of sequence could be placed and the orientation of some scaffolds was not well-supported. The v3 assembly was accompanied by structural annotations of protein-coding genes, transposable elements (TEs) and tandem repeats, and certain RNA genes (Sections III and IV).

Table 5.1. Assembly metrics for *Chlamydomonas* chromosome-level nuclear genome assemblies.

Assembly strain/version/date	CC-503 v4 2008	CC-503 v5 2012	CC-503 v6 2022	CC-4532 v6 2022	CC-1690 2020
Technology	Sanger	Sanger + 454	PacBio + Illumina	PacBio + Illumina	Nanopore + Illumina
Total length (Mb)	112.3	111.1	111.5	114.0	111.1
Unplaced scaffolds/contigs	71	37	42	40	1*
Unplaced length (Mb)	9.68	2.20	1.45	1.72	1.65
Contigs	2,739	1,495	145	120	21
Contig N50 (Mb)	0.09	0.22	2.92	2.65	3.58
GC (%)	64.1	64.1	64.1	64.1	64.1
Gaps/Ns (%)	7.54	3.65	1.66	0.81	<0.01
Transposable elements (%)	9.84	10.61	10.80	12.42	11.24
Microsatellites (%)	1.32	1.43	1.72	1.76	1.65
Satellite DNA (%)	3.33	3.68	4.79	5.25	5.09

Unplaced sequence in v4 and v5 was assembled on scaffolds, while unplaced sequence in all other assemblies was assembled on contigs. Microsatellite refers to tandem repeats with monomer lengths <10 bp, and satellite DNA to those with monomer lengths >=10 bp.

*One unplaced contig forms the right arm of chromosome 15.

Continued efforts yielded two chromosome-level assemblies, reviewed by Blaby et al. (2014). The first of these, v4 (see metrics in Table 5.1), incorporated targeted Sanger sequencing of gaps and repetitive regions. Alongside substantial improvements to contiguity, the assembly length was notably reduced to 112.3 Mb by the removal of redundant regions. The most long-standing assembly version to date, v5, utilized both Sanger and short-read, or next-generation, sequencing. The v5 assembly was 111.1 Mb, with 3.7% gaps and 37 unplaced scaffolds spanning 2.2 Mb (Table 5.1). The contig N50 more than doubled to ~220 kb and over 1,000 gaps were filled relative to v4. As we are writing, the v5 assembly is still the reference, and can be accessed from Phytozome (https://phytozome.jgi.doe.gov/). Given its longevity and that its release coincided with the revolution in omics technologies spurred by the development of Illumina sequencing, v5 was extensively used across an array of analyses including transcriptomics, epigenetics, proteomics. However, many genes still contained sequence gaps (Tulin and Cross 2016), and at least two studies highlighted genes (*PSY1* and *MTHI1*) that were assembled on an incorrect chromosome based on genetic mapping data (Salomé and Merchant 2019; Ozawa et al. 2020). These inconsistencies were not present in v4, indicating that some of the improvements in contiguity realized in v5 had potentially come at the cost of misassemblies.

B. Version 6: Long reads, misassembly corrections and a new reference strain

The development of long-read or third generation sequencing on the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT, also known as “Nanopore”) platforms has enabled the sequencing of reads tens or even hundreds of kb in length. Since these lengths exceed those of most of the repeats in the *Chlamydomonas* genome, these technologies brought substantial improvements to the assembly. Indeed, prior to the v6 genome project, PacBio sequencing was applied to both unicellular and multicellular relatives of *Chlamydomonas*, yielding assemblies more contiguous than that of v5 (Hamaji et al. 2018; Craig et al. 2021a; Yamamoto et al. 2021). Most notably, O'Donnell et al. (2020) produced an assembly of the widely studied CC-1690 (21 gr) in which the 17 chromosomes were represented by only 21 contigs (Table 5.1). This was achieved using a Nanopore dataset sequenced by Liu et al. (2019), which included ultra-long reads that bridged all but the most complex parts of the genome.

The v6 genome project, described in Craig et al. (2022), set out to fill the remaining gaps in the *Chlamydomonas* v5 assembly using a standard approach of high coverage PacBio sequencing for initial assembly, with additional Illumina reads for “polishing”, which refers to the correction of errors introduced by the error-prone long-reads. For the first time, two strains were targeted for assembly: CC-503, the long-term *mt*+ reference strain, and CC-4532, an *mt*– laboratory strain. The provenance of CC-4532 was unclear (Gallaher et al. 2015), although analysis of haplotype blocks (Section V) and identification of shared strain-specific TE insertions suggests that CC-4532 is a subclone of CC-621 (NO^{-}), a strain selected by the Goodenough lab for its high mating efficiency. CC-4532 was initially chosen to

assemble the mating-type *minus* allele of laboratory strains, since only a partially complete sequence of this biologically significant region (Section IV) was available from the divergent Minnesota field isolate CC-2290 (S1 D2) (Ferris et al. 2010). However, since the CC-503 genome was found to carry major structural aberrations (see below), the CC-4532 genome was promoted to be the reference v6 assembly.

Both the CC-503 and CC-4532 PacBio datasets were assembled *de novo*, yielding contigs with N50s >2.6 Mb, an order of magnitude longer than for v5 (Table 5.1). These long contiguous sequences were then manually arranged on chromosomes by mapping to the near-complete CC-1690 assembly. The CC-4532 contigs were syntenic with CC-1690, and Figure 5.1A shows a representation of the resulting CC-4532 v6 chromosomes. This assembly contains only 63 gaps and almost all the v5 unplaced scaffolds are assembled to chromosomes (salmon-colored chromosomal regions in Figure 5.1), with the 40 unplaced contigs in CC-4532 v6 mostly featuring highly repetitive sequences that were not present in v5. Importantly, the majority of filled gaps fall within genes, and almost half contain some novel exonic sequence, as revealed by the CC-4532 v6.1 annotation (Section III). The genic filled gaps are frequently associated with intronic tandem repeats rather than TEs, suggesting that the intron-rich genome architecture of *Chlamydomonas* (Section III) had precluded a more complete assembly based on Sanger and short-read sequencing. The remaining gaps are mostly associated with known large repeats, including centromeres and subtelomeres (Section VI), as well as complex arrays of satellite DNA, so it is expected that almost all genic sequence is represented in v6. Note that the length of CC-4532 v6 (114.0 Mb) is slightly greater than that in past versions, partly due to the rapid expansion of a particular TE during the laboratory culture of CC-4532 (Section V).

As expected, the *de novo* v6 assemblies revealed several misassemblies in v5, which affected 11 chromosomes and ranged from minor local changes to reorganizations of megabases of sequence within and between chromosomes. In Figure 5.1A, colored bands on the CC-4532 v6 chromosomes (other than salmon, see above) mark regions that were reassigned to a different chromosome. Misassemblies in v5 were frequently associated with highly repetitive regions (e.g. centromeres and subtelomeres). As an example, Figure 5.1B shows how sequences from chromosome 15, the most affected one, were scattered over several chromosomes and unplaced scaffolds in v5. This chromosome is distinctively repetitive and gene poor, with a repeat content of 47%, rising to 67% in a ~3.2 Mb internal region where gene content is just 11%. The unique features of chromosome 15 almost certainly explained its past misassembly, and it remains the most fragmented chromosome in the v6 assemblies. Interested readers are directed to the v6 genome paper for details of all corrections (Craig et al. 2022).

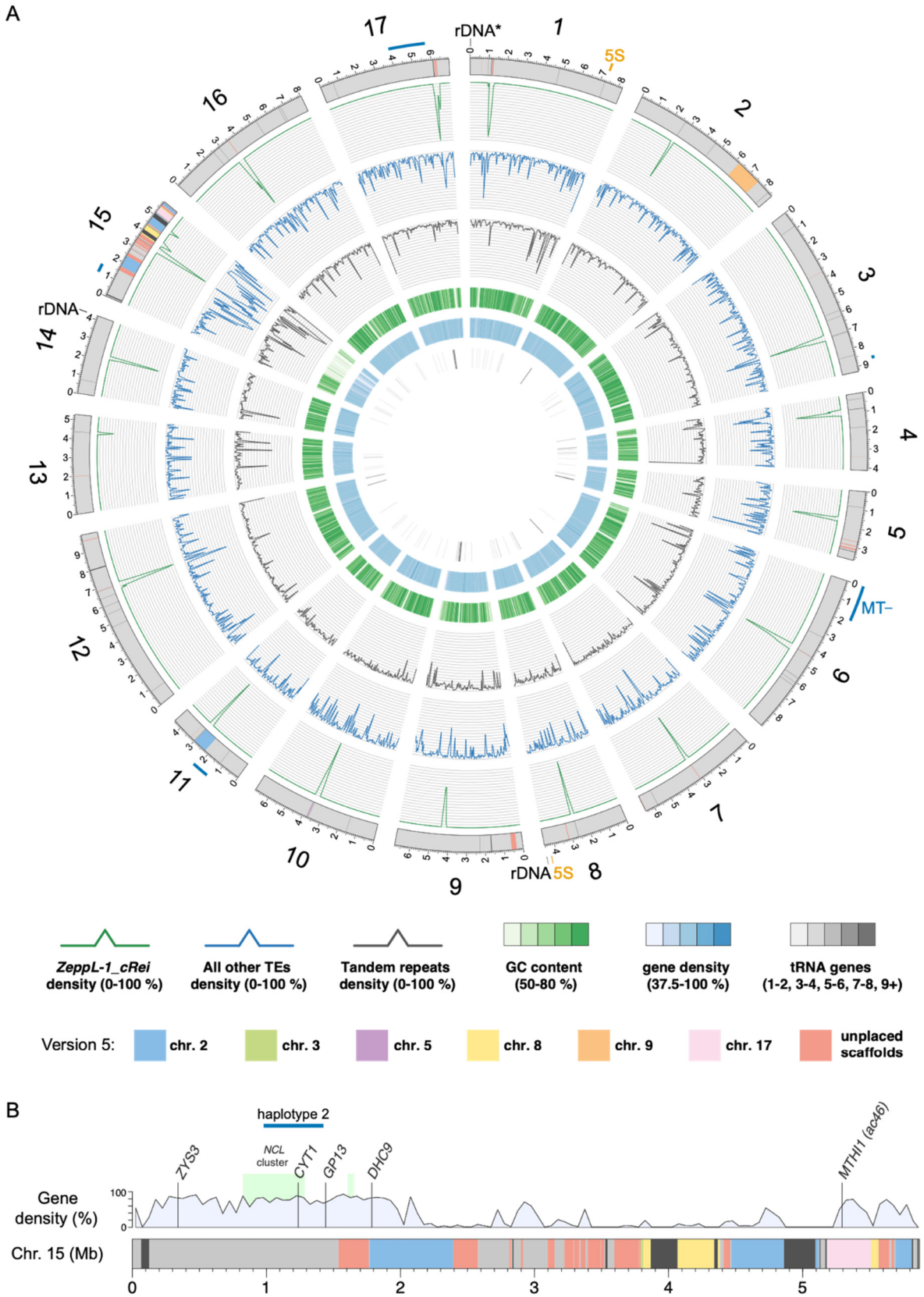
A
rDNA*
5S
MT–
rDNA
rDNA 5S
ZeppL-1_cRei density (0-100 %)
All other TEs density (0-100 %)
Tandem repeats density (0-100 %)
GC content (50-80 %)
gene density (37.5-100 %)
tRNA genes (1-2, 3-4, 5-6, 7-8, 9+)
Version 5:
chr. 2
chr. 3
chr. 5
chr. 8
chr. 9
chr. 17
unplaced scaffolds
B
haplotype 2
ZYS3
NCL cluster
CYT1
GP13
DHC9
MTHI1 (ac46)
Gene density (%)
Chr. 15 (Mb)

Figure 5.1. Overview of the CC-4532 version 6 assembly.
(A) Circos plot (Krzywinski et al. 2009) representation of CC-4532 v6. Grey outer blocks represent chromosomes, with additional colors highlighting genomic regions that in v5 were assembled on other chromosomes or unplaced scaffolds. Dark grey represents gaps between contigs, with any gaps <10 kb increased to 10 kb to aid visualization. All metrics were calculated for 50 kb non-overlapping windows. rDNA and 5S rDNA arrays are marked, with the truncated rDNA array on the left arm of chromosome 1 depicted with an asterisk (see Section V). Dark blue lines exterior to the main plot mark regions where CC-4532 and CC-503 carry alternative haplotypes (e.g. the region featuring the mating-type locus, marked MT^-, on the left arm of chromosome 6).
(B) Linear representation of chromosome 15. Marker genes are from Kathir et al. (2003) and the light green boxes represent the clusters of tandemly repeated and rapidly evolving OPR-RAP (*NCL*) genes identified by Boulouis et al. (2015).

The CC-1690 assembly, which served as a backbone for the v6 assemblies, was further validated via the re-analysis of meiotic recombination data (Figure 5.2) from both molecular mapping (Kathir et al. 2003) and the re-sequencing of tetrad progeny (Liu et al. 2018). Maps based on each dataset showed numerous inconsistencies against the v5 assembly (Figure 5.2A, D). In contrast, only two markers from Kathir et al. (2003), GP332 and *ODA16*, were inconsistent with CC-1690 (Figure 5.2C, F). These regions received unambiguous assembly support and the markers were thus likely historically misplaced. Furthermore, the linkage data from tetrad progeny (Liu et al. 2018) were entirely consistent with CC-1690 (Figure 5.2E, F). While CC-1690 and CC-4532 v6 are congruent, there remained one outstanding inconsistency in CC-503 v6 involving chromosomes 2 and 9 (Figure 5.2B, F). Inconsistencies between these chromosomes in v4 and v5 had been highlighted previously (Lin et al. 2013), and detailed analysis of CC-503 v6 revealed a putative complex reciprocal translocation, with a partial inversion of the sequence translocated from chromosome 9 to 2 (Figure 5.3B). This major chromosomal aberration was confirmed to be unique to CC-503 by inspection of Illumina data from CC-125, its progenitor. The translocation was also present in v5, even though misassembled (Figure 5.3A). Indeed, although its exact assembly details changed between versions, it can be documented as early as v2, implying that the mutation occurred prior to the start of the genome project.

Following this discovery, Craig et al. (2022) curated additional structural mutations by comparing the three available assemblies, with variants found in only one of the strains scored as mutations. Remarkably, more than 70 structural mutations (not including TE insertions), predicted to affect over 100 genes, were found in CC-503 v6. Most are deletions, which cumulatively result in the loss of >300 kb of sequence. Furthermore, approximately two-thirds of the identified mutations are absent in past versions, implying that they occurred during laboratory culture between the original Sanger sequencing of the early genome project and the v6 PacBio sequencing. The most striking of these was a large inversion/deletion double mutation, involving a ~508 kb inversion on the left arm of chromosome 16 and a 47 kb deletion within the inverted sequence (Figure 5.3C). Presumably some of the CC-503 mutations common to all versions occurred under the original mutagenesis of the strain. One such mutation is potentially the cause of the genomic instability: a second deleted region in chromosome 16 contains a RecQ helicase (Cre16.g801898) homologous to *A. thaliana RECQ3* and human *RECQL5*. RecQ helicases play major roles in DNA stability and repair,

including in double-strand break repair pathways (Lu and Davis 2021). Another large deletion entirely removed a prolyl-4 hydroxylase gene (Cre01.g800047). Because prolyl-4 hydroxylases are involved in the formation of the hydroxyproline rich glycoproteins that constitute the *Chlamydomonas* cell wall (Woessner and Goodenough 1994; Keskiaho et al. 2007), this could represent the *cw92* mutation responsible for the cell wall-less phenotype (although segregation of the *cw* phenotype in crosses suggests that there may be more than one causal mutation (Davies 1972; Hyams and Davies 1972)).

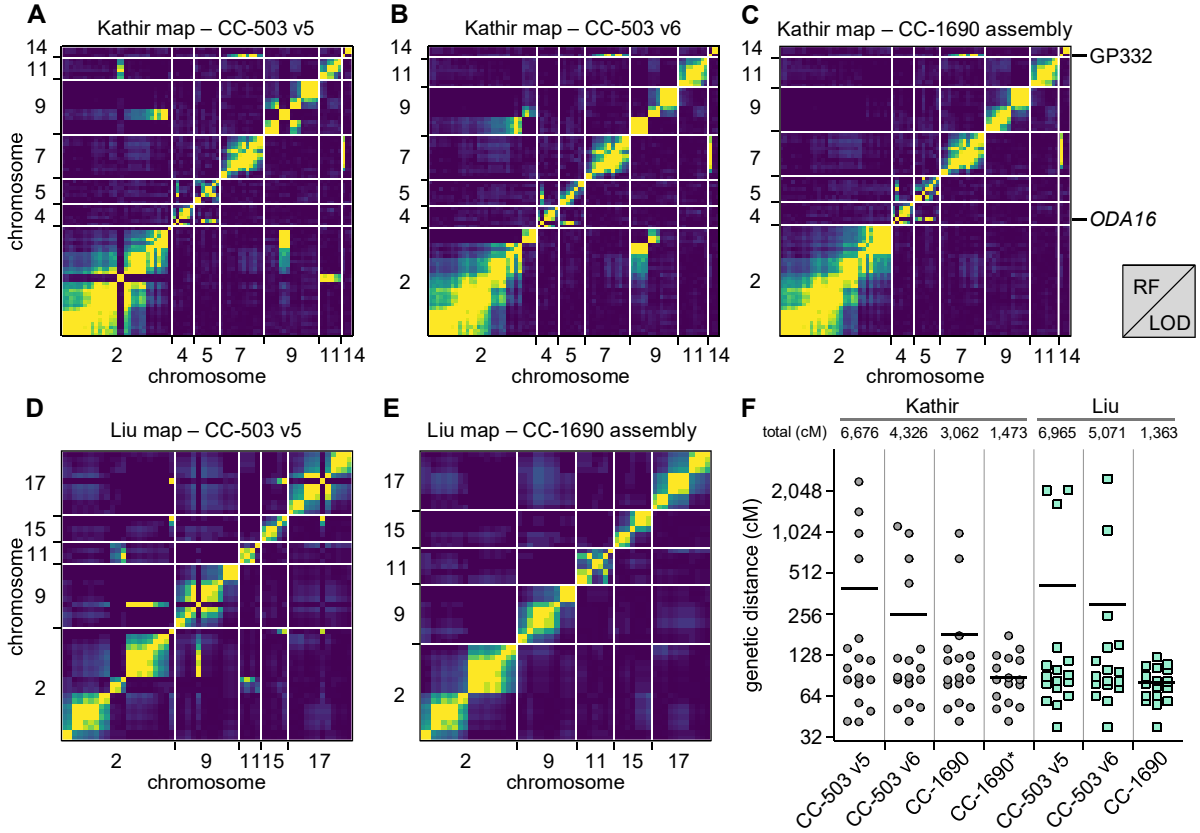

Figure 5.2. Validation of the CC-503 v6 and CC-1690 genome assemblies by recombination maps.

(A) Partial plot of recombination frequencies between molecular markers used by Kathir et al. (2003) in the v5 assembly. Strong linkage is indicated by a yellow color; absence of linkage is shown as dark blue. Seven representative chromosomes are shown.

(B, C) Partial recombination frequency plots between the same molecular markers with updated genomic coordinates according to the CC-503 v6 **(B)** or CC-1690 **(C)** assembly. Note that the markers GP332 and *ODA16* are consistently mis-mapped, as they show strong linkage to a chromosome distinct from the one they were assigned to. Also note that the CC-503 v6 assembly exhibits a remaining inconsistency between chromosomes 2 and 9.

(D, E) Partial recombination frequency plots between informative single nucleotide polymorphisms extracted from Liu et al. (2018), when using the genomic coordinates from the v5 **(D)** or CC-1690 **(E)** assemblies. RF: recombination fraction. LOD: logarithm of the odds. Five representative chromosomes are shown.

(F) Gradual improvement of the estimation of genetic map length, from v5, to CC-503 v6, to CC-1690. Chromosome lengths are plotted in cM for each increment of the genetic maps. CC-1690* denotes the use of CC-1690 genomic coordinates after the removal or the GP332 and *ODA16* molecular markers from the analysis. Total map length, in cM, is listed above each dot plot. Horizontal bar: mean.

CC-4532 v6 is by far the most contiguous annotated *Chlamydomonas* assembly to date, and the first expected to be biologically accurate with respect to the ordering and orientation of all sequence. Together with the fact that it carries the identical haplotypes as the highly unstable past reference CC-503 over more than 95% of the genome (Section V), its completeness justifies the switch to CC-4532 as the primary reference strain. However, CC-4532 v6 also carries a small number of gene-disrupting structural mutations of its own, and a far greater number of unique TE insertions (Section V), showing that, due to evolution in the laboratory, no single strain can provide a perfect reference for all studies. One future solution to this problem could be to create a theoretical "ancestral" reference assembly that would remove strain-specific mutations via the comparison of assemblies from multiple strains. Such a reference genome could also include assemblies of the divergent regions (i.e. haplotype blocks) present among laboratory strains.

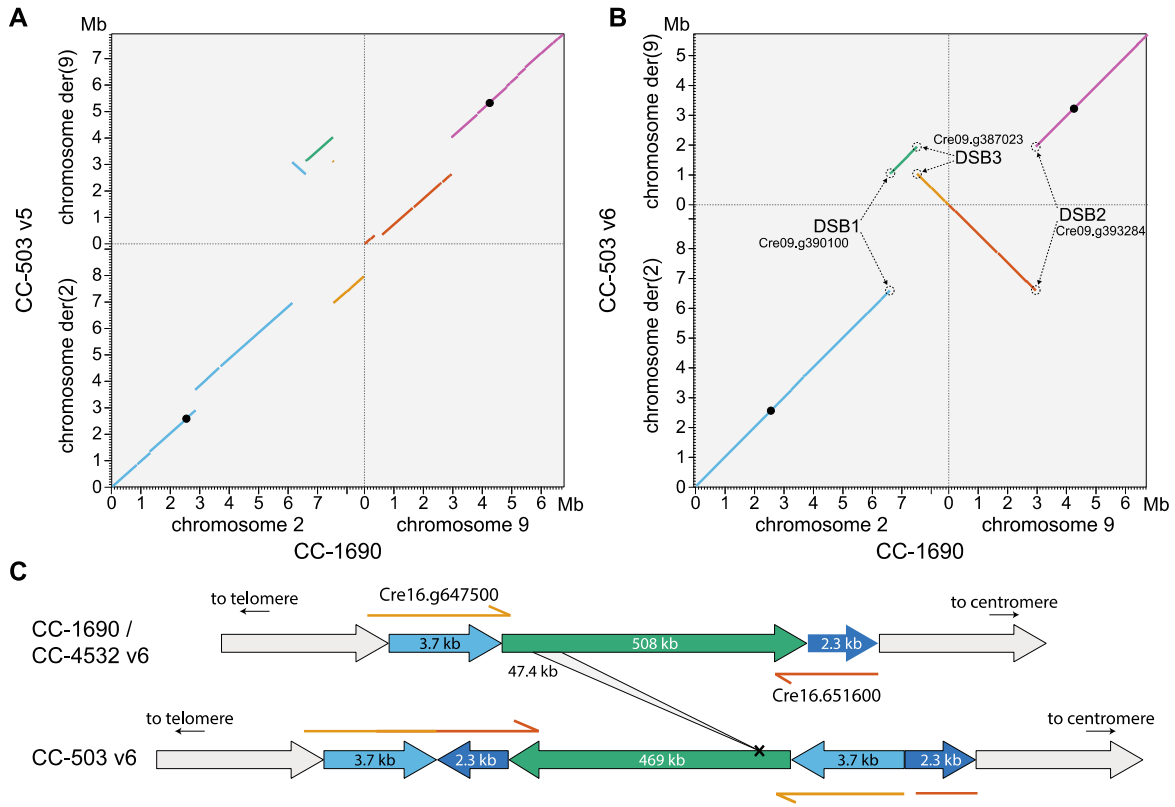

Figure 5.3. Examples of major structural mutations in CC-503
(A, B) Dotplot representations of chromosomes 2 and 9 between v5 (misassembled) and CC-1690 **(A)**, and CC-503 v6 and CC-1690 **(B).** "der" refers to the derived form. Colors represent the major chromosomal fragments involved in the reciprocal translocation and inversion. Double-strand breaks (DSBs) and the genes they disrupt are labelled. Black circles represent putative centromeres. CC-503 chromosomes are named as derivatives (der) based on their centromeres.
(C) Schematic representation of the inversion/deletion double mutation on chromosome 16. The flanks (light and dark blue) of the inversion are duplicated and are shown 50x the scale of the main inverted fragment (green). The 47.4 kb internal deletion is represented by the gray ribbon. The left flank is predicted to have formed a gene fusion in CC-503 v6.1, although this is entirely based on *ab initio* prediction.

III. STRUCTURAL ANNOTATIONS AND GENE ORGANIZATION

A. Versions 3-5: The history of gene model annotation in *Chlamydomonas*

Structural annotations, which define the coordinates of genes and the proteins that they encode, are integral to almost all analyses involving the genome. While this is obviously true for analyses involving protein function (see Vol. 1 Chapter 4), the structural annotations also provide essential information that describes general features of gene organization and genome architecture. Gene models define transcription start sites (TSSs) and terminators, translation start and stop codons, and exon/intron boundaries, and therefore provide annotations for coding sequence, 5' and 3' untranslated regions (UTRs), introns and intergenic sequence. Obtaining high quality structural annotations is fundamental to the functional interpretation of the genome and has been a central goal of each iteration of the *Chlamydomonas* Genome Project.

The structural annotations produced for versions 3-5 of the reference genome have been reviewed by Blaby et al. (2014) and Blaby and Blaby-Haas (2017). The evidence underlying the JGI v3 annotation was primarily based on the Sanger sequencing of almost 200,000 expressed sequence tags (ESTs) primarily from three strains: CC-1690, CC-408 and the highly polymorphic CC-2290 (Asamizu et al. 1999; Asamizu et al. 2000; Shrager et al. 2003; Jain et al. 2007; Merchant et al. 2007). The JGI Annotation Pipeline harnessed this evidence together with protein sequences from model organisms and several gene prediction algorithms, annotating 15,143 protein-coding genes (Table 5.2). Of these, ~60% featured both a start and stop codon, and ~30% had both 5' and 3' UTR sequences (Merchant et al. 2007). An independent annotation effort, GreenGenie and GreenGenie2, annotated 12,387 genes, 78% of which corresponded to genes in JGI v3 (Li et al. 2003; Kwan et al. 2009). Although the intersect between the two gene sets supported the majority of JGI v3 gene models, the proportion of unique genes in each set suggested that further refinement was necessary.

The v4 assembly saw four structural annotation versions released between 2008 and 2012, with each iteration benefiting from extensions to the AUGUSTUS gene prediction algorithm used for annotation (Stanke et al. 2006; Stanke et al. 2008; Specht et al. 2011). The final release, JGI v4.3, featured 17,114 gene models, an increase of almost 2,000 genes compared to JGI v3. Significantly, this update utilized evidence from two emerging sources: >6 million ESTs sequenced on the 454 platform, and homology to genes annotated in the *Volvox carteri* assembly (Prochnik et al. 2010). The v5 annotations were also performed with AUGUSTUS and took full advantage of the transition from ESTs to deep transcriptomic sequencing with RNA-seq. More than one billion RNA-seq reads from an array of experiments were incorporated, including paired-end reads and reads from stranded libraries. The latter were particularly important in enabling the strand of gene models to be determined given the compactness of the genome (see below). Combined with the underlying assembly improvements, the new evidence resulted in many changes to gene models, often involving the splitting and merging of existing genes. Following the JGI v5.3.1 annotation, which

introduced alternative splicing, a final improvement of gene models was released as v5.6, comprising 17,741 genes with 1,785 alternative transcripts. Some locus IDs were changed, stemming from the complication of lifting over loci from v4 (Blaby and Blaby-Haas 2017).

The completeness/quality of an assembly/annotation can be quantified by a BUSCO score, which, in "protein mode", compares a given annotation to a set of proteins that are encoded almost universally by single-copy genes in a specific taxonomic group (i.e. Benchmarking Universal Single-Copy Orthologs). Using a chlorophyte dataset of 1,519 genes, the proportion of missing BUSCOs declined from 3.9% in JGI v3, to 2.0% in JGI v4.3 and 0.8% in JGI v5.6 (Table 5.2). Despite the relative completeness of the v5 releases, several studies highlighted areas of potential improvement in the decade during which they were available. Cross (2015) demonstrated that more than 4,000 gene models had in-frame upstream open reading frames (uORFs), and comparison to *V. carteri* genes showed that many of these were likely to represent genuine N-terminal extensions. As introduced in Section II, Tulin and Cross (2016) revealed that many gaps in the v5 assembly harbored exonic sequences. Blaby and Blaby-Haas (2017) reported that over 100 genes present in JGI v4.3 were not successfully transferred to v5 annotations, including well characterized genes such as *PSBW1*. Using a comparative genomics approach that utilized genomes of the closest relatives of *Chlamydomonas* (Section V), Craig et al. (2021a) identified more than 100 additional high confidence genes that were missing in JGI v5.6. They also found that ~1,000 genes were part of TEs and that several hundred gene models had low coding potential and were unlikely to encode proteins. Finally, Gallaher et al. (2021) used PacBio sequencing of cDNA libraries (i.e. Iso-Seq) to discover polycistronic genes (see below), which were included in an annotation version (v5.7) available only as a supplementary file. The v6 project provided the opportunity to target many of these specific issues, and the Iso-Seq dataset was especially capable of capturing gene models at unprecedented resolution.

Table 5.2. Comparison of protein-coding structural annotations between reference genome versions.

Annotation	JGI v3	JGI v4.3	JGI v5.6	CC-503 v6.1	CC-4532 v6.1
Nuclear genes	15,143	17,114	17,741	16,795	16,801*
Alternative transcripts	82	/	1,789	14,874	14,979
Transposable element genes	/	/	/	647	810
Low coding potential genes	/	/	/	1,435	1,417
Plastome genes	/	/	/	72**	72**
Mitogenome genes	/	/	/	8	8
BUSCO (chlorophyta_odb10, *N*=1,519)	C:92.9% [S:88.9%,D:4.0%] F:3.2%,M:3.9%	C:96.7% [S:96.0%,D:0.7%] F:1.3%,M:2.0%	C:98.9% [S:98.2%,D:0.7%] F:0.3%,M:0.8%	C:100.0% [S:99.3%,D:0.7%] F:0.1%,M:0.0%	C:99.8% [S:98.8%,D:1.0%] F:0.1%,M:0.1%

Note that annotations prior to v6 contained unaccounted TE genes within the main gene set. BUSCO (v4.0) results include the percentage of BUSCOs identified as complete (C), fragmented (F) and missing (M), with complete genes divided into single-copy (S) and duplicated (D) models.

*CC-4532 v6.1 contains 16 *MT+* specific genes that were included on a dedicated *MT+* contig even though their sequences come from CC-503, Craig et al. (2022).
** the three trans-spliced exons of *psaA* are here counted as a single gene.

B. Version 6 structural annotations

As introduced, the improvements seen in the v6 assembly combined with highly informative new evidence provided scope for a substantially improved annotation. Craig et al. (2022) performed *de novo* annotations for both the CC-4532 v6 and CC-503 v6 assemblies, yielding the CC-4532 v6.1 and CC-503 v6.1 gene sets. Approximately 1.6 billion 150 bp stranded and paired-end RNA-seq reads from JGI's CC-1690 Gene Atlas project (https://phytozome-next.jgi.doe.gov/geneatlas/) were incorporated, alongside ~520 million RNA-seq reads from CC-4532, ~6.4 million ESTs and ~1.6 million Iso-Seq reads (Gallaher et al. 2021). Protein homology from 13 green algal gene annotations was also utilized where possible. Genes were predicted using a combination of tools, and the best-scoring model at each locus was retained based on transcriptomic and homology support. The final CC-4532 v6.1 annotation features 16,801 protein-coding genes, with only two missing (0.2%) and one fragmented BUSCOs (Table 5.2). The incorporation of novel exonic sequence in several hundred filled gaps resulted in many changes to gene models including the merger of neighboring gene models previously split by a gap in v5.

The issues highlighted above were also systematically addressed. ORFs were extended at the 5' end where possible, and genes missing in the v5 annotations were added. Twelve genes encoding selenoproteins (Novoselov et al. 2002) were manually curated, all of which had been misannotated due to the use of the canonical stop codon "UGA" to encode selenocysteine, and a small number of other genes were manually corrected (e.g. *LAO1*, which contains a 5 bp exon that was missed by all annotation pipelines). Approximately 1,400 predicted gene models were designated as having low coding potential and were separated from the main annotation. These models are generally very short and their ORFs are poorly conserved both among *Chlamydomonas* field isolates and between species. Some may be long noncoding RNAs (lncRNA), a class of genes that are under-investigated in *Chlamydomonas* (see below). Careful efforts were also applied to separately annotate genes that are part of TEs. Finally, for the first time the organelle genomes and gene annotations were included with the v6 releases, after having been updated based on the latest data (Cavaiuolo et al. 2017; Salinas-Giegé et al. 2017; Gallaher et al. 2018).

C. Gene IDs

In the course of annotating the v4 assembly a unified nomenclature was adopted for *Chlamydomonas* locus names, in the form "**Cre**YY.**g**NNNNNN", where "YY" represents the chromosome (or scaffold) number and "NNNNNN is a numerical identifier unique for each locus, increasing initially by 50 from the start of chromosome_01 (Blaby et al. 2014). This nomenclature is still in use today, ensuring that genes can be conveniently traced across assembly and annotation versions. Novel genes in v6, with no equivalent in v5, were given a

NNNNNN number starting at 800000. As of v6, "_XXXX" will be systematically appended to the locus name to denote (by its CC-number) the strain the sequence comes from. In addition, to account for the substantial changes in chromosome location and to denote the correct order of genes (Section II), an additional "associated locus ID" tag was introduced, in the form "XXXX_YY_NNNNN", where "YY" is the chromosome number and "NNNNN" represents the order of loci in the CC-4532 v6 reference (even numbers for the top strand, odd for the bottom strand). Loci in other assemblies should be given the same "NNNNN" number as their CC-4532 ortholog, whenever applicable. These IDs provide a spatially informative counterpart to the original "Cre" IDs, since the latter may now be misleading due to assembly changes (e.g. a gene named Cre02.gNNNNNN may be on chromosome 9, and so on).

D. Gene structure and organization

Despite its relatively large size, the *Chlamydomonas* genome is highly compact with respect to gene space, with genic sequence spanning more than three-quarters of the genome. The median intergenic distance is 307 bp and ~85% of intergenic tracts are <2 kb (Figure 5.4A). The remaining 15% comprise ~80% of total intergenic sequence and are repeat-rich (~66% repeats, compared to ~16% for the short tracts) (Table 5.3). The inclusion of Iso-Seq data in the v6 annotations surprisingly revealed that the longest isoforms of ~29% of neighboring genes overlap, emphasizing the compactness of gene organization. Furthermore, 72% of intergenic regions separate genes on opposite strands, indicating a strong bias in favor of the convergent/divergent orientation of successive transcription units. Thus, many transcripts can base-pair via their 3' UTRs, a phenomenon that has been shown to impact gene expression in yeast (Sinturel et al. 2015). *Chlamydomonas* 3' UTRs are also relatively long, with the median length of 679 bp considerably longer than that of *A. thaliana* (244 bp) or *D. melanogaster* (187 bp), for example.

Some *Chlamydomonas* promoters have been studied in great detail, e.g. those of *HSP70A*, *CYC6*, *PSAD1* and *NIT1/NIA1*, and the role of the TATA box has been outlined (Quinn and Merchant 1995; Fischer and Rochaix 2001; Loppes and Radoux 2002; Lodha et al. 2008). However, TSSs have not been mapped precisely genome-wide, and a comprehensive description of *Chlamydomonas* promoters is still lacking. Although Iso-Seq data are expected to improve this situation, TSSs would ideally be analyzed using dedicated methods that target capped mRNAs. Nonetheless, putatively active promoter regions have also been characterized epigenetically. Using chromatin immunoprecipitation sequencing (ChIP-seq), Ngan et al. (2015) identified four chromatin states indicative of promoter regions, which featured a progressive combination of the active marks H3K4me3, H3K27ac, H3K9me3 and H3K36me3 (where K is a specific lysine on histone H3, me3 is trimethylation and ac is acetylation). H3K4me3 has since been demonstrated to be a highly stable epigenetic mark for *Chlamydomonas* TSSs (Strenkert et al. 2021). Several studies have also identified marks associated with transcriptionally silenced chromatin (Vol. 1 Chapter 6). Furthermore, Fu et al. (2015) characterized adenine methylation (N^6-methyldeoxyadenosine, 6mA) at promoters. 6mA is centered on AT dinucleotides and forms a bimodal distribution with peaks either side

of the TSS. The 6mA enrichment within the peaks shows a periodicity of 130-140 bp, which corresponds to linker regions between adjacent nucleosomes. The TSS bimodal distribution was observed in more than 80% of genes and was generally associated with active transcription and higher gene expression. Downstream of promoter regions, a Kozak-like sequence, similar to that of humans, which flanks more efficient initiators/start codons, has been identified in *Chlamydomonas* (Cross 2015). Genetic manipulation of the Kozak-like sequence has been demonstrated to influence translation *in vitro* (Gallaher et al. 2021).

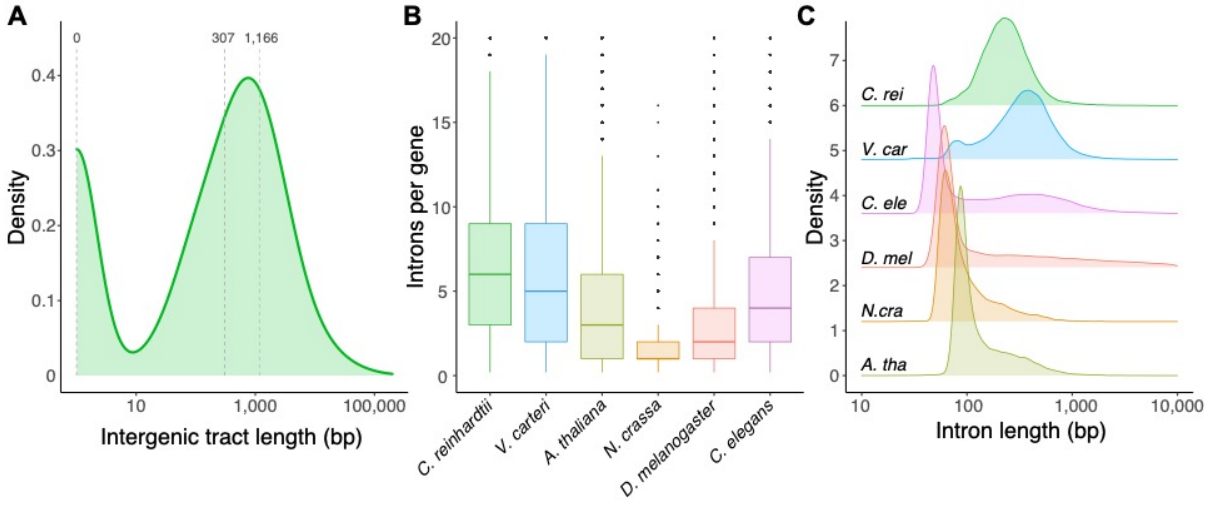

Figure 5.4. Noncoding sequence in the Chlamydomonas genome.
(A) Density plot of intergenic tract lengths. Dotted lines mark the first, second and third quartiles. Note that lengths are plotted on a log scale, and tract lengths of 0 bp were arbitrarily set to 1 bp for plotting.
(B) The number of introns per gene, per species.
(D) The distribution of intron lengths per species.

Table 5.3. Genomic metrics of *Chlamydomonas* site classes.

Site class	Subdivision	Mb	Genome (%)	GC (%)	TE (%)	Microsatellite (%)	Satellite (%)
CDS		38.06	33.65	70.05	0.40	0.91	3.70
	0D	24.53	21.69	64.03			
	2D	5.14	4.54	84.78			
	4D	8.23	7.28	78.87			
5' UTR		4.01	3.54	54.56	3.70	0.21	1.26
3' UTR		10.20	9.02	58.43	7.04	0.41	1.31
introns		34.70	30.68	62.11	5.00	3.48	4.92
intergenic		26.14	23.11	61.78	43.65	1.53	10.27
	≤2 kb	5.45	4.81	59.97	13.50	0.98	1.47
	>2 kb	20.70	18.30	62.26	51.58	1.67	12.58

CDS refers to coding sequence. 0D, 2D and 4D refer to site degeneracy within codons (i.e. zero-fold, two-fold and four-fold degenerate sites), where 0D sites are nonsynonymous and result in amino acid changes and 2D and 4D are either partially or entirely synonymous. Metrics are given for all

intergenic tracts, as well as for tracts split into those shorter or longer than 2 kb. Calculated using the longest isoform per gene only.

In contrast, the sites where transcripts end are more easily deduced from transcriptomic data due to the poly-A tail (Vol. 1 Chapter 6). The main signal governing cleavage and polyadenylation, UGUAA, was first described by Silflow et al. (1985). Zhao et al. (2014b) analyzed polyadenylated transcripts from Sanger, Illumina and 454 datasets, confirming the UGUAA polyadenylation signal (later refined to UGUAAC by Li and Du (2014)), as well as a tendency for polyadenylation to start after a genome-encoded adenine. They proposed a high level of alternative polyadenylation, including in coding sequences and introns, a finding not supported by the Poly(A) Tag Sequencing (PAT-Seq) approach of Bell et al. (2016). In the latter study, ~25% of genes had more than five distinct polyadenylation sites, but they usually clustered less than 24 bp apart. Forty percent of genes showed two or more clusters, but clusters were rare outside of 3' UTRs and no link was found between alternative polyadenylation and differential expression. Their data can be browsed at PlantAPAdb (http://www.bmibig.cn/plantAPAdb/APAcatalog.php) (Zhu et al. 2020). (See also Vol. 2 Chapter 6).

E. Intron abundance and lengths

The length and abundance of introns are major contributing factors to the high gene density of the *Chlamydomonas* genome. Intronic sequences comprise almost a third of the genome and each gene contains eight introns on average, which is similar to human genes and substantially higher than most model organisms with comparable genome sizes (Figure 5.4B). The intron richness is shared by *V. carteri* and all unicellular and multicellular close relatives of *Chlamydomonas* that have been sequenced and annotated thus far (Craig et al. 2021a). In a study modelling intron evolution across eukaryota, Csuros et al. (2011) inferred that a major expansion of introns occurred in early chlorophyte evolution, with high intron densities maintained in certain taxa by balanced rates of intron loss and gain. Although it has not yet been explored analytically, one possible explanation for the retention of introns in *Chlamydomonas* is the prevalence of non-homologous end-joining (NHEJ) over homologous recombination (HR) in the repair of double-strand breaks (Zorin et al. 2005; Ferenczi et al. 2021). The relative roles of these alternative pathways are thought to govern intron evolution, with HR predicted to drive intron deletion, and NHEJ implicated in both the acquisition and loss of introns (Farlow et al. 2011).

Chlamydomonas introns also tend to be unusually long. The median intron length is 230 bp, considerably longer than the median exon length of 151 bp. Indeed, the short introns of 60-110 bp that are typically dominant in small eukaryotic genomes such as *A. thaliana* and *D. melanogaster* constitute only ~5% of introns in *Chlamydomonas* (Figure 5.4C). Once again, this feature of gene organization is shared by close relatives of *Chlamydomonas*. Merchant et al. (2007) suggested that intron length may have been driven by repeat expansion. However, TEs are underrepresented in introns, especially introns located towards the 5’ of genes (Philippsen et al. 2016), and the total TE density of introns is comparable to that of UTRs

(Table 5.3). In contrast, introns are enriched for tandem repeats (Zhao et al. (2014a), Table 5.3), although the overall intronic repeat content remains relatively low at ~13%.

An alternative explanation of intron length would be if introns harbored a substantial number of functional sites, i.e. specific sequences with a biological function. For example, Kang and Mitchell (1998) described an enhancer within the first intron of the 5' UTR of the dynein gene *DIC2*, and Croft et al. (2007) documented thiamine riboswitches in introns of *THI4* and *THIC1*. Craig et al. (2021a) found that ~19% of intronic sites intersected with conserved elements, genomic regions inferred to be functionally constrained between species. Baier et al. (2020) demonstrated that the presence of several different *Chlamydomonas* introns can greatly improve transgene expression, suggesting that introns could be retained by selection. Although greater upregulation was observed for introns located closer to the TSS, the insertion of additional introns had an additive effect. In two cases, they attempted to identify regulatory elements involved in intron-mediated enhancement, but did not find any specific intronic regions associated with the regulatory effect. Indeed, upregulation could be driven by several introns taken from other species, provided the splice sites were altered to match the *Chlamydomonas* consensus (G/GTGAG … CACAG/G). They did however find that the removal of regions flanking the splice junctions impaired efficient splicing. This supports the results of Raj-Kumar et al. (2017), who identified a putative branch point motif and additional G-rich and C-rich sequences within 50 bp of the 5' and 3' splice sites, respectively, that may function as intronic splice enhancers. Although there are undoubtedly regulatory elements present within introns, as well as other functional sequences (e.g. introns involved in alternative splicing or containing noncoding RNA genes, see below), these studies suggest that functional constraint cannot solely explain intron length.

Overall, it seems likely that the gene expression machinery of *Chlamydomonas* is well adapted to genes with many introns of moderate lengths, although the mechanism underlying the association between introns and gene expression remains unclear.

F. Alternative splicing

Alternative splicing (AS) has been documented in detail for only a few algal genomes, but in *Chlamydomonas* the accumulation of transcriptome data has led to a steady rise in the number of annotated alternative transcripts (Vol. 1 Chapter 6). Initially based on small genomic sequences (Li et al. 2003) or comparison with assembled ESTs (Jain et al. 2007), AS was later analyzed on successive versions of the genome in conjunction with ever expanding transcriptome datasets (Labadorf et al. 2010; Raj-Kumar et al. 2017; Pandey et al. 2020). The latter study focused on AS events that show diurnal rhythmicity and proposed an accessory role in regulating gene expression. The v6 annotations integrated thousands of alternative transcripts (Table 5.2), but it must be stressed that they have not been confirmed by proteomic studies at the time of writing. By comparison to the "main" transcript (with .1 suffix, ideally the most abundant, but often simply the longest), AS events can be categorized as intron retention (failure to excise an intron), alternative donor or acceptor (use of a different 5'- or 3'- splice site, respectively), exon retention or exon skipping. Based on the

results of Raj-Kumar et al. (2017), this is the order of their frequency in *Chlamydomonas* (from highest to lowest). This would tend to indicate that the major mode of splicing is by "intron definition", whereby the spliceosome assembles on the splice sites of a given intron, rather than on exons as in organisms with very long introns. Nonetheless, the intron-richness of the genome may also suggest the presence of exonic-splice enhancers, which are bound by serine/arginine-rich proteins and enhance splicing at neighboring intron-exon boundaries (Blencowe 2000). Exonic splice enhancers have not yet been studied in *Chlamydomonas*, although a potential example was reported by Lin et al. (2018b). Finally, alternative transcription (use of a different TSS) or alternative polyadenylation (see above) can also generate transcript variants.

The depth at which the transcriptome is being explored now allows the identification of even rare transcript variants, some of which undoubtedly simply reflect the ambiguity of the sequence signals used by the transcription, splicing and cleavage/polyadenylation machineries. What fraction of this diversity is biologically meaningful, i.e. produces alternative proteins or mediates regulation of gene expression, remains uncertain. The consequences of AS have been fully documented only for a handful of genes, including *CGE1* (Schroda et al. 2001), *ANK22* (Li et al. 2003), *CTH1* (Moseley et al. 2002), *CCM1* (Fukuzawa et al. 2001) and the long 5′ extension described in *FLU1* (Falciatore et al. 2005). The thiamine-PP riboswitches inside introns of the *THI4* and *THIC* genes are examples where regulated AS leads to the formation of unstable transcripts or non-functional proteins, extinguishing gene expression (Croft et al. 2007). AS seems poorly conserved between *Chlamydomonas* and *Volvox* (Kianianmomeni et al. 2014), but comparison with more closely related species or between *Chlamydomonas* strains could reveal cases where AS seems to be selected.

G. Polycistronic gene expression

An exciting discovery made possible by Iso-Seq data was that of polycistronic gene expression. Gallaher et al. (2021) used the long cDNA reads in combination with H3K4me3 ChIP-seq (marking promoters, see above), proteomics and *in vitro* experimental validation to identify 87 polycistronic loci, most of which were dicistronic. Among these was *REX1*, described many years earlier by Cenkci et al. (2003) as a single mRNA encoding two proteins, REX1S and REX1B, involved in DNA repair. Polycistronic genes share a single promoter and poly(A) tail, and as expected are highly co-expressed in RNA-seq datasets. Although this work focused on obligately polycistronic genes, a similar number of facultatively polycistronic loci were observed. Many of these cases also possessed a single promoter, meaning that the upstream gene could be transcribed independently but the downstream gene was only transcribed as part of a polycistronic transcript. The ORFs of polycistronic genes were shorter than those of monocistronic genes, and the inter-ORF distances were substantially less than those between neighboring monocistronic genes. Polycistronic loci were also detected in other chlorophytes, and they may be an ancient feature of green algal genomes. Although a mechanism has not yet been characterized, the patterns observed were most consistent with leaky ribosome scanning, in which the first ORF

is bypassed by the ribosome at a certain frequency in favor of the downstream ORF. If valid, this explanation may go some way to explaining the prevalence of uORFs in *Chlamydomonas* genes (Cross 2015), which are present in thousands of genes even after accounting for the in-frame cases that were previously misannotated (see above). While the presence of multiple uORFs in isoforms may inhibit protein synthesis (Moseley et al. 2002), their abundance in the *Chlamydomonas* transcriptome suggests that ribosomes can translate ORFs downstream of alternative start codons.

H. Non-coding RNA genes

Intron splicing is performed by the spliceosome, a complex machinery that assembles progressively on the donor and acceptor splice sites around conserved small U-rich RNAs. The *Chlamydomonas* spliceosomal small nuclear RNAs (snRNAs) have been characterized by Jakab et al. (1992), Kis et al. (1993), and Jakab et al. (1997), who studied their base-paring properties and showed that U1, U2, U4 and U5 are transcribed by Pol II, while U6 is transcribed by Pol III. Table 5.4 lists the snRNA genes in v6. As noted by Merchant et al. (2007), some linkage can be observed between genes of same type (U1 on chr 1 and 6, U2 on chr 9, U4 on chr 2, U5 on chr 6, U6 on chr 8). Many snRNA genes were located within introns of protein-coding genes but could be transcribed from their own promoter. Interestingly, several U1 and U2 loci give rise to long, spliced and polyadenylated transcripts sharing the downstream exons with their host gene. It is therefore possible that their expression requires processing from a polyadenylated precursor. The multi-subunit POL II-associated Integrator complex, which in animals governs transcription and maturation of snRNAs, is partially conserved as the "DSP complex" in *A. thaliana* (Liu et al. 2016) but is absent in *Chlamydomonas* and most green algae.

In eukaryotes, the 18S, 5.8S and 28S ribosomal RNA (rRNA) genes are co-transcribed by Pol I from a ribosomal DNA (rDNA) element present in many copies in a few large arrays in the genome. Similarly, the Pol III-transcribed 5S rRNAs are also very often clustered (Haeusler and Engelke 2006). *Chlamydomonas* is no exception (Table 5.4). Two forms of 5S rRNA were sequenced by Darlix and Rochaix (1981), form I being more abundant than form II. In the v6 genome, they map to two clusters of 5S genes, found respectively on chromosomes 1 (69 genes, form II) and 8 (14 genes, form I). The clusters are composed of, respectively, 14 and 3 tandem arrays, each comprising 2 to18 copies of the 5S rDNA (120 nt long) separated by a spacer sequence (respectively 622 and 520-530 nt) conserved within a cluster but widely divergent between the clusters. The better conservation of 5S, spacer and inter-array sequences within than between clusters suggests evolution by concerted birth-and-death mechanisms as described in other species (Pinhal et al. 2011). A complex history of array expansion and duplication has also likely led to the truncation of the 5' and 3' genes in each array. The CC-4532 genome thus encodes 41 5S pseudogenes, including 6 with an internal deletion. Surprisingly, the 5S rDNA loci on chromosome 8, more heavily expressed based on Darlix and Rochaix (1981), are found within introns of RNAseq-supported but probably non-coding RNA, transcribed from the other strand. Whether these unusual lncRNAs play a role in the transcription of the 5S rDNA in *Chlamydomonas* remains to be established.

About 350 copies of the large rDNA element (18S, 5.8S and 28S) have been found in the genome (Howell 1972; Chaux-Jukic et al. 2021), located in two large clusters at the right ends of chromosomes 8 and 14, forming the subtelomeric region as is the case in *A. thaliana* and many other species. In addition a short degenerate and probably non-functional array is found at the left end of chromosome 1 (Chaux-Jukic et al. 2021). Interestingly, this truncated array was found to be a peculiarity of laboratory strains, since field isolates and closely related species appear to carry a third intact array on chromosome 1. In all cases, the rDNA is transcribed towards the telomere, but the telomere-proximal copies are hypermethylated, and hence are probably transcriptionally silent (Chaux-Jukic et al. 2021).

Small nucleolar RNA (snoRNA) genes are responsible for post-transcriptional modifications of the non-coding RNAs. They were analyzed in the original genome paper, and this was refined by Chen et al. (2008) who identified 322 snoRNA genes. They could be mapped to 320 loci on the CC-4532 v6 assembly, but the actual number of snoRNA genes might be higher. *Chlamydomonas* snoRNAs are grouped into 118 families, split into 74 box C/D (guiding 96 site-specific RNA 2'-O-ribose methylations) and 44 box H/ACA (guiding 60 pseudouridylation events). Three quarters of the snoRNA genes were arranged in 76 clusters, most of them apparently created by local tandem duplications. More than 90% of both singleton and clustered snoRNAs were located within protein-coding gene models encoded on the same strand, and usually within or overlapping with introns. It is therefore advisable to examine whether the phenotype ascribed to the disruption of such protein-coding genes could be due to the loss of a snoRNA family. Table 5.4 also lists the very abundant U3 snoRNA (Antal et al. 2000) involved in the initial cleavage of the rRNA precursor and encoded by a single gene (previously included in the now deprecated Cre07.g350976), as well as the 7S cytosolic RNA of the Signal Recognition Particle.

Chlamydomonas transfer RNA (tRNA) have been described by Merchant et al. (2007) and Cognat et al. (2008) who annotated 259 tRNA genes in the v3 genome. They commented on their propensity to form clusters originating from duplication, as well as their intron richness (~60% of tRNAs contained an intron). Although absent from subsequent annotations, tRNAs were re-annotated in GFF format for CC-4532 and CC-503 v6 and are visible as a browser track. This annotation was integrated into the tRNA database (http://plantrna.ibmp.cnrs.fr/), which also incorporates data on tRNA modification enzymes and amino-acyl tRNA synthetases. The CC-4532 v6 assembly contains 325 tRNA genes, including the tRNA[Ser]Sec that allows incorporation of selenocysteine at UGA codons (Rao et al. 2003). Heavy clustering (>5 genes less than 2.5 kb apart) is observed on chromosomes 5, 9, 12 and 17 (Figure 5.1A).

The long non-coding RNAs of *Chlamydomonas* have thus far received little attention, but the development of long read transcriptomic analysis should soon allow their full characterization. Using Illumina sequencing, Li et al. (2016) have already identified 1,440 candidate lncRNA genes, 65% of which are intergenic and about half are single exon. Strenkert et al. (2021) validated ~30% of these models by identifying enrichment of

H3K4me3 at their TSSs. They also found that candidate lncRNA genes associated with H3K4me3 enrichment were more highly expressed than those lacking enrichment. As noted above, some lncRNAs may participate in the expression of 5S rRNA, and a role in subtelomere organization has been proposed by Chaux-Jukic et al. (2021), but their wider role remains unknown. Finally, several studies have identified small RNA genes (e.g. micro RNAs and small-interfering RNAs) in *Chlamydomonas* (Zhao et al. 2007; Valli et al. 2016; Müller et al. 2020), which play multiple roles in gene regulation and silencing.

Table 5.4: Chlamydomonas RNA genes

name	description	accession	length	#loci	transcribed by	notes	clusters (# loci)	position / protein genes*
U1	snRNA, spliceosomal	X70869.1	164	6	Pol II	initiates splicing by base pairing with the 5′ splice site	chr_01 (2); chr_06 (2)	I:3; i:2; 3U:1
U2	snRNA, spliceosomal	X71483.1	192	7 + 1Ψ	Pol II	base-pairs with the branch point sequence	chr_09 (2)	I:1; i:4; 3U:1
U4	snRNA, spliceosomal	X71485.1	138	2	Pol II	forms a duplex with U6 to regulate its activity	chr_02 (2)	I:2
U5	snRNA, spliceosomal	X67000.1	111	7	Pol II	forms a triple-snRNP with U4 and U6 and base-pairs with the 5' exon		I:3; i:2; 3U:4; d:1;a:2
U6	snRNA, spliceosomal	X71486.1	101	13	Pol III	binds to U4, then to U2 to form the catalytic center	chr_08 (3); chr_08 (6)	I:13
5S	cytosolic 5S rRNA	X02706, X02707	122	83 + 40Ψ	Pol III	two forms, I more abundant than II	chr_01 (69+30Ψ); chr_08 (14+10Ψ)	I: chr1 ai: chr_08
45S	cytosolic 18S, 5.8S and 25S rRNA	KX781334 EU410621		~350	Pol I	precursor 45S rRNA is cleaved to form the RNA components of the ribosome	subtelomeric at 8_R, 14_R and 1_L**	I
7SL	cytoplasmic RNA of Signal Recognition Particle	X71484.1	290	2	Pol III	promotes hydrolysis of GTP, releasing SRP from its receptor and the ribosome		I:1; 3U:1
U3	box C/D snoRNA	AJ001179	523	1	Pol III	involved in 45S rRNA cleavage		I:1
CrACAn	Other box C/D snoRNA	EU410622 to EU410808	128-205	187	Pol II	74 families guiding methylation at 96 rRNA and 3 U6 sites	80 singletons + 242 in 76 clusters	i: >90% (2-7 per protein-coding gene)
CrCDn	H/ACA snoRNA	EU410809 to EU410943	67-135	135	Pol II	44 families guiding pseudouridylation at 60 rRNA and 2 U6 sites		
trnX	transfer RNAs	-	71-129	325	Pol III	~60% contain an intron	several clusters on chr 05, 09, 12 and 17	I: all

* I: intergenic; 5U: 5'-UTR; i: intronic; 3U: 3'-UTR; d: dicistronic; a: antisense

**_L: left end; _R: right end

IV. GENOME ARCHITECTURE

A. Centromeres

Although the approximate location of centromeres could be inferred from linkage data (Preuss and Mets 2002), the genomic coordinates and sequence characteristics of centromeric regions have only recently been reported. Lin et al. (2018a) found that regions known to be centromere-linked on 15 chromosomes are characterized by stretches of sequence spanning several hundred kb that feature multiple genes encoding reverse transcriptase. These regions behaved as centromeres in meiotic tetrads. Craig et al. (2021a) further assessed these regions, reporting that most of the reverse transcriptase domains are encoded by multiple copies of a specific *L1* long interspersed nuclear element (LINE), *L1-1_CR* (Kapitonov and Jurka 2004), which is otherwise entirely absent from the genome. A subsequent phylogenetic analysis demonstrated that *L1-1_CR* is more closely related to *Zepp*, the major centromeric component in the trebouxiophyte alga *Coccomyxa subellipsoidea* (Blanc et al. 2012), than to any other *L1* families in *Chlamydomonas*. *Zepp* elements also form one cluster per chromosome in *C. subellipsoidea*, which likely results from a nested insertion mechanism in which new copies insert within existing copies, creating tandem arrays of mostly 5' truncated elements (Higashiyama et al. 1997). Homologs of *L1-1_CR*, which were collectively referred to as *Zepp*-like (*ZeppL*) elements, were identified in similar genomic clusters in several unicellular and multicellular close relatives, although none were detected in *V. carteri*. *L1-1_CR* was given the synonym *ZeppL-1_cRei* to distinguish it from other *L1* elements in the genome.

As mentioned in Section II, the repetitiveness of the putative centromeric regions was often responsible for past misassemblies. After correcting these issues, each chromosome features a single highly localized *ZeppL* cluster, except for chromosome 15 where two additional minor clusters (<30 kb) are observed (Figure 5.1A). Based on the CC-1690 assembly, in which all putative centromeres are entirely assembled except for chromosome 15, the clusters range from 51 kb to 320 kb, with a mean length of 192 kb. More than 95% of centromeric sequence is assembled in CC-4532 v6, although the putative centromeres of only five chromosomes are assembled without gaps.

Chlamydomonas centromeres can therefore be classed as "transposon-rich" and predominantly based on a single TE family, similar to the centromeres in *Dictyostelium discoideum* (Glöckner and Heidel 2009), for example. Nonetheless, many outstanding questions remain. Length aside, it is not yet clear if major compositional differences exist between the *ZeppL* clusters of different chromosomes. The clusters are also enriched for other TEs (Figure 5.1A), with *ZeppL* elements constituting ~60% of the total sequence. Collectively, the clusters account for ~25% of total TE sequence despite spanning only ~3% of the genome. Pericentromeric regions have not yet been defined. Although satellite DNA does not appear to be a major component of the *ZeppL* clusters, satellite arrays are found flanking certain clusters e.g. the satellite *MSAT-2_CR* forms a boundary between the *ZeppL* elements and the right arm of chromosomes 11 and 13 (Chaux-Jukic et al. 2021), and other

satellite-rich regions are observed on different chromosomes (e.g. 4 and 5). Characterizing the localization of the centromeric histone H3 (CenH3) will be an essential step in investigating what exact regions within the clusters function as centromeres (as recently performed, for example, in *D. melanogaster* (Chang et al. 2019)). Finally, it remains to be seen whether *ZeppL* elements form the centromeres of other distantly related green algae, given their presence in *Chlamydomonas* and its closest relatives, as well as the distantly related *C. subellipsoidea.*

B. Telomeres and subtelomeres

In *Chlamydomonas* and its close relatives (including *C. incerta* and *C. schloesseri*, Section V), the telomere consists of an 8-nt (TTTTAGGG)n repeat, with one additional T compared to other green algae and most plants (Petracek et al. 1990; Fulnečková et al. 2012). Eberhard et al. (2019) demonstrated that the median length of the telomere varies between strains from ~300 bp to ~1 kb, and that a fraction carry blunt ends as opposed to the usual 3' single-stranded extension. As in most eukaryotes, a telomerase reverse transcriptase (*CrTERT*) is involved in their maintenance.

While three of the chromosome ends carry rDNA arrays in the subtelomeric region (Section III), the subtelomeric regions of the other 31 present a unique organization. They consist of large arrays of an ~850 bp satellite repeat called *Sultan* (for SUbtelomeric Long TANdem repeats), followed by a conserved *Spacer* sequence, where transcription is initiated in the direction of the centromere (Figure 5.5). This presumably non-coding transcript is composed of a conserved exon, a long first intron characterized by G-rich microsatellite sequences, and variable downstream exon(s). Each chromosome end carries a specific *Sultan*, repeated between 1 and 46 times, usually transitioning in phase into the telomere via a TTTAGG sequence at the end of the *Sultan*. Similarity is strongest between copies of the *Sultan* monomer within the same subtelomere, indicating preferential local tandem duplication. Indeed, occasional expansions and contractions of the *Sultan* monomer within the same subtelomere have been observed during long-term culture in the laboratory (López-Cortegano et al. 2022). The *Sultan* arrays are marked by CG methylation, which together with various types of TEs usually found downstream of the *Spacer* constitute a large heterochromatic region, as often found at subtelomeres.

Analysis of assemblies and Illumina datasets shows that the *Sultan* sequences, and even the number of copies, are generally well-conserved among laboratory strains, but can vary substantially in field isolates. For example, the genome of the highly divergent North Carolina isolate CC-2931 showed non-cognate *Sultan* arrays at a majority of subtelomeres. In addition, it lacked the *Suber* and *Subtile*, two other types of repeats found in a subset of the subtelomeres of laboratory strains between the telomere and *Sultan* array (class C in Figure 5.5, see Chaux-Jukic et al. (2021) for details). Finally, although a similar organization of tandem repeats was observed at the subtelomeres of close relatives, *Sultan* itself is so far unique to *C. reinhardtii*, speaking to the rapid evolution of these regions. This comprehensive

description of subtelomeres was only made possible by the extremely contiguous Nanopore-based assembly of CC-1690.

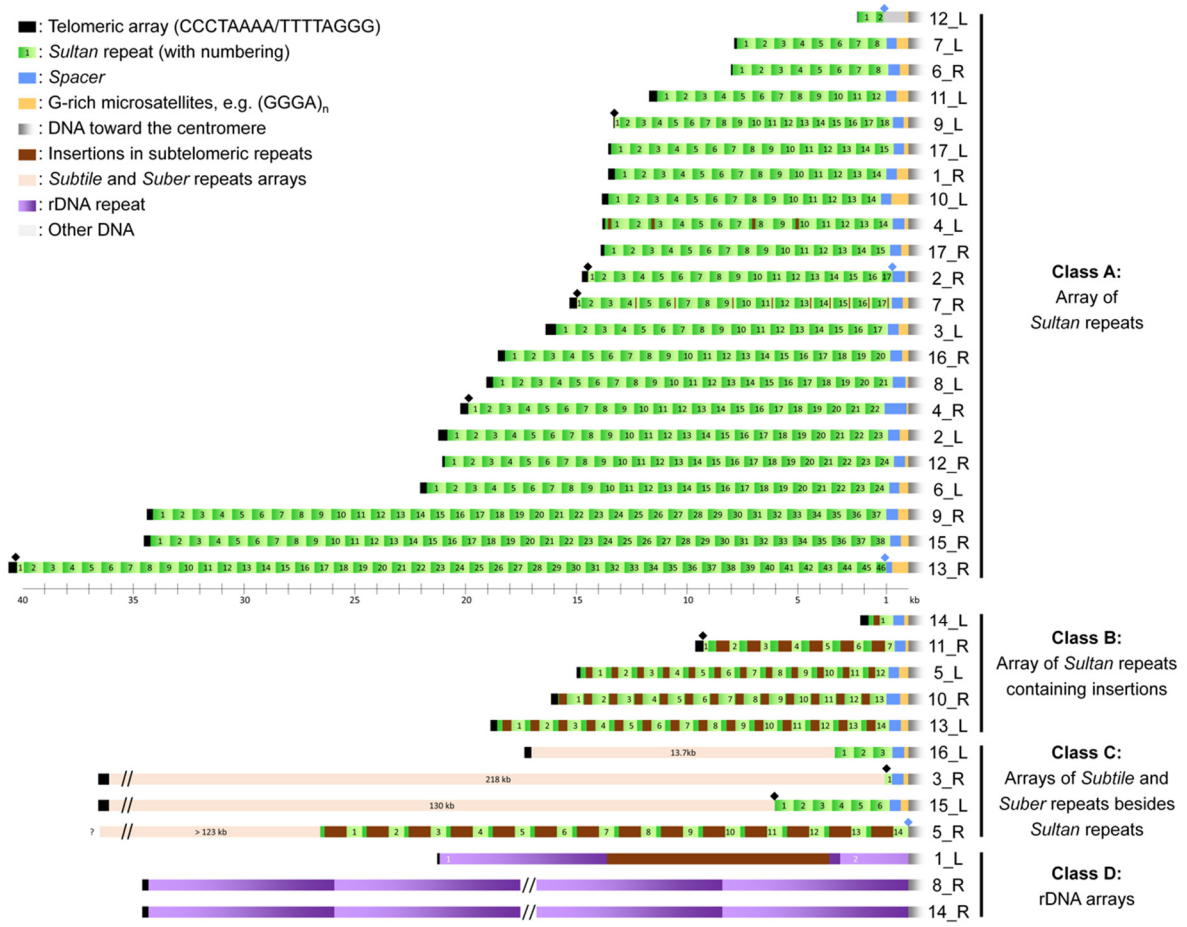

Figure 5.5. Architecture of subtelomeres in *Chlamydomonas* strain CC-1690.
Left and right ends (_L and _R, respectively) of CC-1690 chromosomes are depicted with telomeres on the left-hand side. Class A subtelomeres comprise a telomere tract (black), a tandem array of *Sultan* repeats (green; numbering starts on the telomere side), a *Spacer* sequence (blue) and a G-rich microsatellite (yellow). Distinct large DNA insertions (brown) found in the *Sultan* repeats define the class B subtelomeres. Class C subtelomeres contain repeats of the *Suber* and *Subtile* elements (in pink) upstream of the *Sultan* array. Arrays of rDNA (purple) compose class D subtelomeres. From Chaux-Jukic et al. (2021).

C. Transposable elements

Chlamydomonas has had a rich but understated role in TE research. Given its phylogenetic distance from other models, *Chlamydomonas* TEs have often been among the first representatives of entirely new TE clades. In the pre-genome era, a small number of active TEs were experimentally characterized. Day et al. (1988) and Day and Rochaix (1991) described *TOC1*, an active 5.7 kb retrotransposon that features split terminal repeats unlike any other TE described at the time, and in general does not produce target site duplications (TSDs) upon insertion. Ferris (1989) described *Gulliver*, an active 12.2 kb DNA transposon

characterized by 15 bp terminal inverted repeats (TIRs) and 8 bp TSDs. *TOC1* and *Gulliver* have since been used as analytical models of transposition (e.g. Wu-Scharf et al. (2000), Casas-Mollano et al. (2008)). A handful of other active TEs followed: the DNA transposons *Tcr1* (Schnell and Lefebvre 1993; Ferris et al. 1996) and *Tcr3* (Wang et al. 1998), and a second unusual retrotransposon *Pioneer1* (Graham et al. 1995). The nonautonomous DNA transposon *TOC2* was identified as an insertion polymorphism between a laboratory strain and CC-2290 (Day 1995). Later experimental work characterized the *Gypsy* long terminal repeat (LTR) element *CrREM1* (Perez-Alegre et al. 2005), the non-autonomous DNA transposon *Bill* and the non-autonomous retrotransposon *MRC1* (Kim et al. 2006).

The availability of the early assembly versions enabled researchers to directly curate repetitive sequences of interest. Although nonautonomous, *TOC1* was linked to TEs (e.g. *TOC3*) that encoded proteins with reverse transcriptase and tyrosine recombinase domains, placing these elements within the emerging *Dictyostelium* intermediate repeat sequence-like (DIRS) group (Goodwin and Poulter 2004). *Pioneer1* was also classified as a DIRS element (Goodwin and Poulter 2001; Poulter and Goodwin 2005). Kojima and Fujiwara (2005) described the novel LINE clade *Dualen*, which unlike all other LINEs encode both restriction-like endonuclease and apurinic/apyrimidinic endonuclease-like endonuclease. Cognat et al. (2008) annotated short interspersed nuclear elements (SINEs), which are nonautonomous elements derived from tRNA genes that in *Chlamydomonas* mostly rely on *Dualen* elements for their activity. A thorough annotation effort by Kapitonov and Jurka yielded the vast majority of the 119 *Chlamydomonas* TEs deposited in Repbase (https://www.girinst.org/repbase/). This led to the description of *Novosib*, a new superfamily of DNA transposons (Kapitonov and Jurka 2008; Yuan and Wessler 2011), while other annotations were improved; for example *Gulliver* was classified in the *hAT* superfamily based on the identification of its transposase (Kapitonov and Jurka 2006). The v4 and v5 assemblies continued to be sources of biological novelty, as demonstrated by the contribution of *Chlamydomonas* sequences to the discovery of the *Helitron2* clade of Helitron DNA elements (Bao and Jurka 2013), the *Kyakuja-Dileera-Zisupton* (*KDZ*) superfamily of DNA transposons (Böhne et al. 2012; Iyer et al. 2014), and to *Chlamys*, a major new clade of *Penelope*-like elements (Craig et al. 2021b).

Although the Repbase library served as a long-standing reference set, almost all of the TE consensus models were curated from the v3 assembly and the completeness of the library was unclear. Aiming to update the library, Craig (2021) performed exhaustive manual curation of both existing and novel TEs using the v5 and v6 assemblies. This effort more than doubled the number of TE consensus sequences to a total of 269, resulting in an ~50% increase in TE sequence identified genome-wide. Many of the existing models were extended or improved, and the classification of several TEs were either extended or entirely changed. For example, the DNA transposons *Tcr1* and *Tcr3* were classified to the *Kyakuja* and *EnSpm* superfamilies, respectively. All previously annotated TEs were systematically given synonyms in the updated library to distinguish the original and updated consensus sequences, and in some cases to correct past misclassifications (see Table 5.5 for examples).

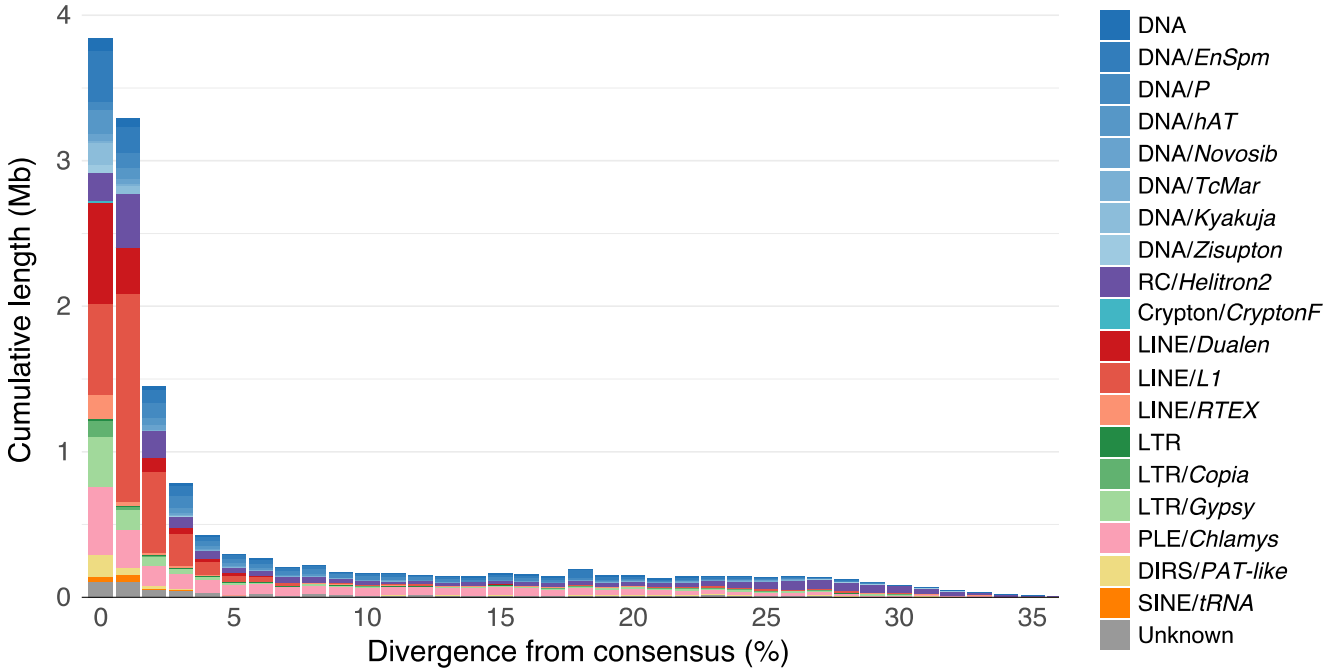

Figure 5.6. *Chlamydomonas* transposable element landscape.
Cumulative length of TE sequence plotted against divergence from TE family consensus sequences. TE sequence is colored by order and superfamily (if known). DNA = DNA transposon, RC = rolling circle and PLE = *Penelope*-like element, all other acronyms are introduced in the main text. TE abundance and divergence estimates were generated against the CC-1690 assembly.

Almost 80% of TE sequence is found in intergenic space, with regions particularly enriched for TEs including the centromeres, subtelomeres (downstream of the *Sultan* arrays, see above) and much of chromosome 15. Interestingly, there is a negative relationship between chromosome length and TE content, which remains even after excluding chromosome 15 (Pearson's r = -0.53). The correlation is weakened but still strongly significant if centromeres are removed, and generally the smallest chromosomes (particularly 4 and 5) are more repetitive (Figure 5.1A).

Two other results are particularly noteworthy concerning *Chlamydomonas* TEs. First, the species contains an astounding diversity of TEs, with families from eight of the nine described eukaryotic orders present, representing 16 superfamilies (see Wells and Feschotte (2020) for order and superfamily definitions). As a point of comparison, the *A. thaliana* genome contains TEs from five orders and 12 superfamilies. Second, copies from the same TE family exhibit very low divergence, with 80% of TE copies exhibiting <5% divergence from their family consensus sequence (Figure 5.6). This implies that most TE families were either recently active or are currently active, since older inactive copies would be expected to randomly accumulate mutations. Furthermore, it implies that inactive TEs are efficiently purged, speaking to the compactness of the genome. These results have several implications for genome evolution and host TE defense, which is thought to act at both the transcriptional

and post-transcriptional level in *Chlamydomonas* via a number of partly independent mechanisms (van Dijk et al. 2006).

Combining experimental and genomic observations (Section V), active transposition has been characterised for ~15 TE families in laboratory strains (Table 5.5). Interestingly, a similar number of active families have recently been characterized in the field isolate CC-2931, although there is no overlap between the active families in laboratory strains, while only three families appear to be active in CC-1952 (López-Cortegano et al. 2022). This suggests that the repertoire and number of active TEs may be highly heterogenous among isolates, in line with past observations that *Gulliver* is absent from several field isolates (Ferris 1989) and that *Pioneer1* is absent from laboratory strains (Graham et al. 1995).

Table 5.5. Active transposable elements in laboratory strains.

Class	Order	Superfamily	(sub)Family	Craig (2021) synonym	Action	Reference
retrotransposon	LINE	*L1*	*L1-2_CR*	*L1-2_cRei*	autonomous	Craig et al. (2022)
retrotransposon	LINE	*L1*	*L1-6_CR*	*L1-3_cRei*	autonomous	Craig et al. (2022)
retrotransposon	LTR	*Gypsy*	*Gypsy-8_CR*	*Gypsy-7a_cRei*	autonomous	Craig et al. (2022)
retrotransposon	LTR	*Gypsy*	*CrREM1*	*Gypsy-19_cRei*	autonomous	Perez-Alegre *et al*. (2005)
retrotransposon	PLE	*Chlamys*	*NonLTR-5_CR*	*Chlamys-3_cRei*	autonomous	Craig et al. (2022)
retrotransposon	PLE	*Chlamys*	*MRC1*	*Chlamys-N8_cRei*	nonautonomous (*Chlamys-8_cRei*)	Kim *et al*. (2005)
retrotransposon	DIRS	*PAT-like*	*TOC1*	*PAT-N1a_cRei*	nonautonomous (*PAT-3_cRei*)	Day *et al*. (1988); Day & Rochaix (1991)
retrotransposon	?	?	/	*unknown-4b_cRei*	nonautonomous (*L1-6_CR*)	Craig et al. (2022)
DNA	TIR	*hAT*	*Gulliver*	*hAT-1_cRei*	autonomous	Ferris (1989)
DNA	TIR	*hAT*	/	*hAT-4_cRei*	autonomous	Craig et al. (2022)
DNA	TIR	*hAT*	*hAT-N2_CR*	*hAT-N2_cRei*	nonautonomous (*Gulliver*)	Craig et al. (2022)
DNA	TIR	*hAT*	*Bill*	*hAT-N9_cRei*	nonautonomous (*hAT-4_cRei*)	Kim *et al.* (2005)
DNA	TIR	*Novosib*	*hAT-N4_CR*	*Novosib-N4_cRei*	nonautonomous (*Novosib-4_cRei*)	Craig et al. (2022)
DNA	TIR	*EnSpm*	*Tcr3*	*EnSpm-1_cRei*	autonomous	Wang *et al*. 1998
DNA	TIR	*Kyakuja (KDZ)*	*Tcr1*	*Kyakuja-1_cRei*	autonomous	Schnell & Lefebvre (1993)
DNA	TIR	*Kyakuja (KDZ)*	/	*Kyakuja-3_cRei*	autonomous	Craig et al. (2022)

The citation provided is for the first reported activity of the family. For nonautonomous families the hypothesized autonomous partner is listed in parentheses. *TOC1* putatively relies on *PAT-3_cRei* for

its transposition, which was originally named *TOC2* by Goodwin and Poulter (2004); however, this name was already taken for an unrelated DNA transposon (Day 1995). For families newly described by Craig (2021) no original name is given. Note that some TEs may be active only in specific laboratory strains.

D. Genome-wide patterns of methylation

Both cytosine and adenine methylation have been documented in *Chlamydomonas* (see Vol 1. Chapter 6). As described in Section III, adenine methylation (6mA) occurs in a highly specific context at promoters (Fu et al. 2015). Cytosine methylation (specifically C^5-methylcytosine, or 5mC) of the nuclear genome has been estimated to occur at low levels, ~1-5% for CG sites, and ~0.25-2.5% for CHG and CHH sites (Feng et al. 2010; Lopez et al. 2015). In contrast to plants, CHG and CHH methylation is not targeted to TEs and other repeats, and instead appears to be uniformly distributed across chromosomes with enrichment in exons (Feng et al. 2010). Conversely, CG methylation shows a slight enrichment in gene bodies and a far more substantial enrichment in repeats. Lopez et al. (2015) identified 23 highly repetitive loci where CG methylation reached 80%. Subsequent reassessment of these hypermethylated regions showed that most coincided with the putative centromeres. Further analysis of 5mC methylation using Nanopore reads, which are mappable in far more repetitive regions than bisulfite short-read data, confirmed the hypermethylation of centromeres and revealed additional hypermethylation in subtelomeres (see above) and some other regions of high repeat density (Chaux-Jukic et al. 2021; Craig et al. 2022). Although experimental support is variable (Lopez et al. 2015), these results suggest a role for CG methylation in gene silencing, as may also be the case in *V. carteri* (Babinger et al. 2007).

Recently, the novel base modification C^5-glyceryl-methylcytosine (5gmC) was discovered in *Chlamydomonas* (Xue et al. 2019). Although present at only ~1,000 sites genome-wide, 5gmC appears to be an intermediate in a novel 5mC demethylation pathway catalyzed by TET/JBP (ten-eleven translocation/J-binding protein) enzymes (Aravind et al. 2019). One *Chlamydomonas* TET-JBP gene, *CMD1*, was characterized by Xue et al. (2019). Genome-wide 5mC was doubled in a *cmd1* mutant, resulting in downregulation of several genes. However, *CMD1*, and indeed most of the 12 TET-JBP genes in the reference genome, are part of various TEs, and due to TE copy number variation the number of TET-JBP genes is variable between strains (Aravind et al. 2019; Craig et al. 2022). This situation mirrors that observed in several fungi (Iyer et al. 2014), and it may be the case that these TEs selfishly induce demethylation to regulate their own expression and activity. The impact of TET-JBP enzymes on gene expression across the general transcriptome remains unclear.

E. Base composition, mutation, recombination and codon usage

One of the most striking features of the *Chlamydomonas* genome is its overall GC content of 64.1%. GC content is also relatively uniform; if the genome is split into nonoverlapping 20 kb windows, 98% of the genome falls between 58.5% and 69.6% GC. However, considering site classes, variation in GC content is more pronounced: it is highest in coding sequences

(70.1%), lowest in 5' (54.6%) and 3' UTRs (58.4%), and intermediate in intronic (62.1%) and intergenic (61.8%) sequences (Table 5.3). Breaking down coding sequence by site degeneracy, i.e. the number of alternative nucleotides at a position that would not alter protein sequence, GC content is 78.9% at four-fold degenerate (4D) sites, 84.8% at two-fold degenerate (2D) sites and 64.0% at zero-fold degenerate (0D) sites, implying a strong preference for GC at synonymous sites. This suggests the operation of translational selection, whereby particular codons are translationally optimal and are selected for (see Bulmer (1991), Rocha (2004), Plotkin and Kudla (2011)). Naya et al. (2001) defined a set of optimal codons in *Chlamydomonas* and demonstrated that the major trend in codon usage between genes was correlated with gene expression, a pattern characteristic of translational selection. Of the 21 optimal codons, 13 featured a C in the 3rd position, five a G and three a T, while no optimal codons ended in A. Cognat et al. (2008) demonstrated another classic property of translational selection, finding a positive relationship between optimal codons, the tRNA gene copy number and the abundance of the tRNAs needed to decode them. Barahimipour et al. (2015) showed experimentally that optimizing codon usage in a transgene resulted in higher translational efficiency and mRNA stability.

This however does not explain the GC content of noncoding sequence, which is presumably shaped by forces including mutation, GC-biased gene conversion (gBGC) and selection. *Chlamydomonas* is an excellent model for studying mutation properties, and several studies have estimated the mutation rate and spectra in the species (Ness et al. 2012; Sung et al. 2012; Ness et al. 2015; Ness et al. 2016), as well as in the closely related *C. incerta* (López-Cortegano et al. 2021). In the largest study of this type, Ness et al. (2015) estimated an overall mutation rate (μ) of 11.5 x 10^{-10} per site per generation when selection was experimentally kept to a minimum. The mutation rate at C:G sites was 2.4x higher than that at A:T sites, and mutations from C:G to T:A were the most common class of mutation, occurring at a rate almost 2x higher than expected under a balanced spectrum (Figure 5.7). The estimated equilibrium GC content under the inferred mutation spectrum is 29%, implying that gBGC or selection, or both, play major roles in driving increased GC content.

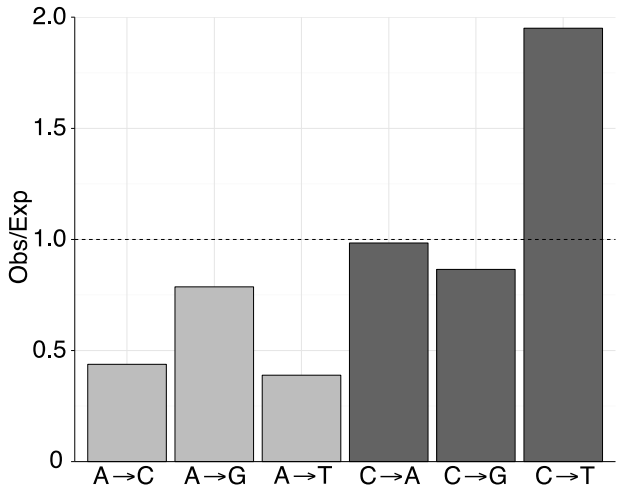

Figure 5.7. *Chlamydomonas* genome-wide mutation base spectrum.
Spectrum was inferred from 5,716 single nucleotide mutations from six *Chlamydomonas* field isolates. Expectation was calculated based on random mutation with respect to each base. Strands are represented symmetrically e.g. C to T signifies C:G to T:A. From Ness et al. (2015).

gBGC is a process in which GC/AT heterozygous sites near the double-strand breaks that initiate recombination are preferentially converted to GC over AT (Duret and Galtier 2009). It can occur at the double-strand breaks of both crossover and non-crossover recombination events. However, establishing its role would require a reliable description of fine-scale variation in recombination rate in *Chlamydomonas*, which is not yet available, even though it would greatly benefit the application of quantitative and population genetics analyses. Liu et al. (2018) performed whole-genome re-sequencing on 21 tetrads from two crosses and observed 24.4 crossovers per tetrad per meiosis, or ~1.4 crossovers per chromosome (equivalent to ~12 cM/Mb). The overall rate of gene conversion from crossovers was 13x higher than non-crossovers, due to crossover recombination both occurring at a higher rate and resulting in longer gene conversion tracts. However, only non-crossover gene conversion events were found to be GC-biased (68.6% bias). The authors found weak but significant positive correlations between both recombination categories and GC content at local scales (10-50 kb), while only the correlation between non-crossovers and GC content was significant at longer distances (100-200 kb). In another approach, Flowers et al. (2015) and Hasan and Ness (2020) each studied variation in genome-wide recombination using whole-genome re-sequencing of field isolates. Flowers et al. (2015) inferred that recombination was reduced towards the ends of chromosome arms, while Hasan and Ness (2020) found that recombination was highest flanking genes within longer intergenic tracts (>2kb) and in coding sequences, and lowest in UTRs. Given the lower frequency of the GC-biased non-crossover events, and that sexual reproduction is likely to be relatively infrequent in the wild (Hasan and Ness 2020), the contribution of gBGC to the GC content of the *Chlamydomonas* genome remains to be demonstrated.

Therefore, selection for increased GC may exist in *Chlamydomonas*. Weissman et al. (2019) developed a hypothesis that linked selection for higher GC content to DNA repair in prokaryotes, noting that species with high GC contents were associated with certain environments that induce higher rates of DNA damage and double-strand breaks (e.g. soil microbes due to desiccation and spore formation). They found a positive association between GC content and the presence of the NHEJ repair pathway, which may be favored in such species relative to the slower HR. They hypothesized that higher GC may increase NHEJ efficiency, providing a selective advantage for GC over AT alleles. Given that *Chlamydomonas* strongly favors NHEJ, it may be that DNA repair has played a role in the evolution of GC content in the species.

F. The mating-type locus

The *Chlamydomonas* mating-type locus is located on the left arm of chromosome 6 and consists of three domains, the ~82 kb T (telomere-proximal), ~204-396 kb R (rearranged) and

~116 kb C (centromere-proximal) domains (De Hoff et al. 2013). The T and C domains are syntenic between *MT+* and *MT–,* while the R domain features several rearrangements and contains the only mating-type-specific genes. The R domain of *MT+* is ~192 kb larger than *MT–* since it contains mating-type-specific autosomal insertions and an ~160 kb tandemly repeated region known as the "16 kb repeats" (De Hoff et al. 2013). Ferris et al. (2010) sequenced the *MT–* locus of CC-2290 to facilitate a direct comparison to the *MT+* locus assembled in the CC-503 reference genome. The CC-4532 v6 assembly now provides an annotated and gapless *MT-* locus, while CC-503 v6 continues to provide an annotated *MT+* assembly, which fortunately does not appear to have been affected by any of the structural mutations typical of the strain (Craig et al. 2022). To facilitate the analysis of data from strains of both mating type, the *MT+* specific regions of the CC-503 *MT+* allele were appended to the CC-4532 v6 assembly as a standalone contig. Note that there is only one ancestral haplotype for each locus among laboratory strains (*MT+* haplotype 1 and *MT–* haplotype 2, see definitions in section V), inherited from the two parents of opposite mating-type.

Crossover recombination is suppressed across the R domain, although the shared genes of *MT+* and *MT–* have not undergone significant differentiation due to non-crossover gene conversion (De Hoff et al. 2013; Hasan et al. 2019). Conversely, mating-type-specific genes have long been known to possess unusual characteristics, and both the *MT–* specific *MID1* and *MT+* specific *FUS1* exhibit very low values for optimal codon usage and low GC content in both coding sequence and introns (Ferris et al. 1996; Ferris and Goodenough 1997). These patterns are in line with the reduced selection efficacy (allowing drift to lower GC) and lack of gBGC resulting from the absence of all recombination (see above). The mating-type loci of volvocine algae are homologous to *Chlamydomonas*, and the region is one of the few that have been studied in fine detail from a comparative perspective (Vol. 1 Chapter 10).

V. GENOME EVOLUTION

A. Genomic variation among laboratory strains

The laboratory strains are a large collection of clonally maintained cultures that are descended from a single diploid zygospore isolated in Massachusetts, 1945. The early history of the strains is complicated and often poorly documented, and traditional models split the strains into three sublines based on the distribution of pairs of opposite mating-type strains to different research groups in the 1950s (Proschold et al. 2005; Harris 2009) (see Vol. 1 Chapter 1). These strains have been maintained as clones in various laboratories and culture centers for approaching 70 years and are sometimes referred to as "wild-type" laboratory strains, although some lines have acquired mutations. Many additional strains have been produced by crosses between "wild-type" laboratory strains and their progeny, including the new reference genome strain, CC-4532.

Gallaher et al. (2015) performed whole-genome Illumina re-sequencing of 39 strains, including both the *mt+* and *mt–* representatives of the wild-type sublines. They identified 41

regions (i.e. haplotype blocks) on 13 of the chromosomes, collectively covering ~25% of the genome, where some strains differ at ~2% of sites. These regions were collectively referred to as “haplotype 2”, meaning that “haplotype 1” was arbitrarily defined relative to CC-503 (and by default the wild-type strain CC-125, or 137c+). For example, CC-4532 carries haplotype 2 in five regions that collectively span 4.6% of the genome (Figure 5.1A), while other haplotype 2 blocks are found in CC-1690. The presence of only two alternative haplotypes is in line with the expectation that the strains are derived from a single zygospore. Gallaher et al. (2015) further found that the classical wild-type strains did not correspond to four hypothetical meiotic products. Instead, five lines were proposed, represented by the strains CC-1691 (Line I), CC-1009 (Line II), CC-1690/CC-1010 (Line III), CC-124 (Line IV) and CC-125 (Line V). This implies that some crosses were made in Smith’s laboratory prior to the distribution of wild-type sublines. A schematic of the haplotype variation among strains, corrected for v5 mis-assemblies, is shown in Figure 5.8.

Despite the high divergence between the haplotypes, it is important not to misinterpret this as evidence that the laboratory strains are unusual in any way. The genomes of any random pair of field isolates sampled from the same location differ at ~2% of sites (Craig et al. 2019). Therefore, the two haplotypes of laboratory strains most likely simply represent the genetic differences between the two parental haploid individuals that once existed in the potato field from which the zygospore was sampled. Nonetheless, this level of genetic diversity is among the highest recorded in eukaryotes (Leffler et al. 2012). As a result, the sheer number of variants that can exist between laboratory strains carrying different haplotype configurations greatly complicates the analysis of *Chlamydomonas* sequencing data. This is particularly true for identifying causative mutations, where one must distinguish ancestral variants that exist between the two haplotypes from those that have occurred in the laboratory by mutation. Computational methods such as described in Lin and Dutcher (2015) greatly help the comparison of sequencings from multiple strains. In addition, Gallaher et al. (2015) proposed sets of primers that enable identification of which haplotype block a strain carries, as well as a computer program that partly reconstructs the genome of a strain based on its haplotype block makeup.

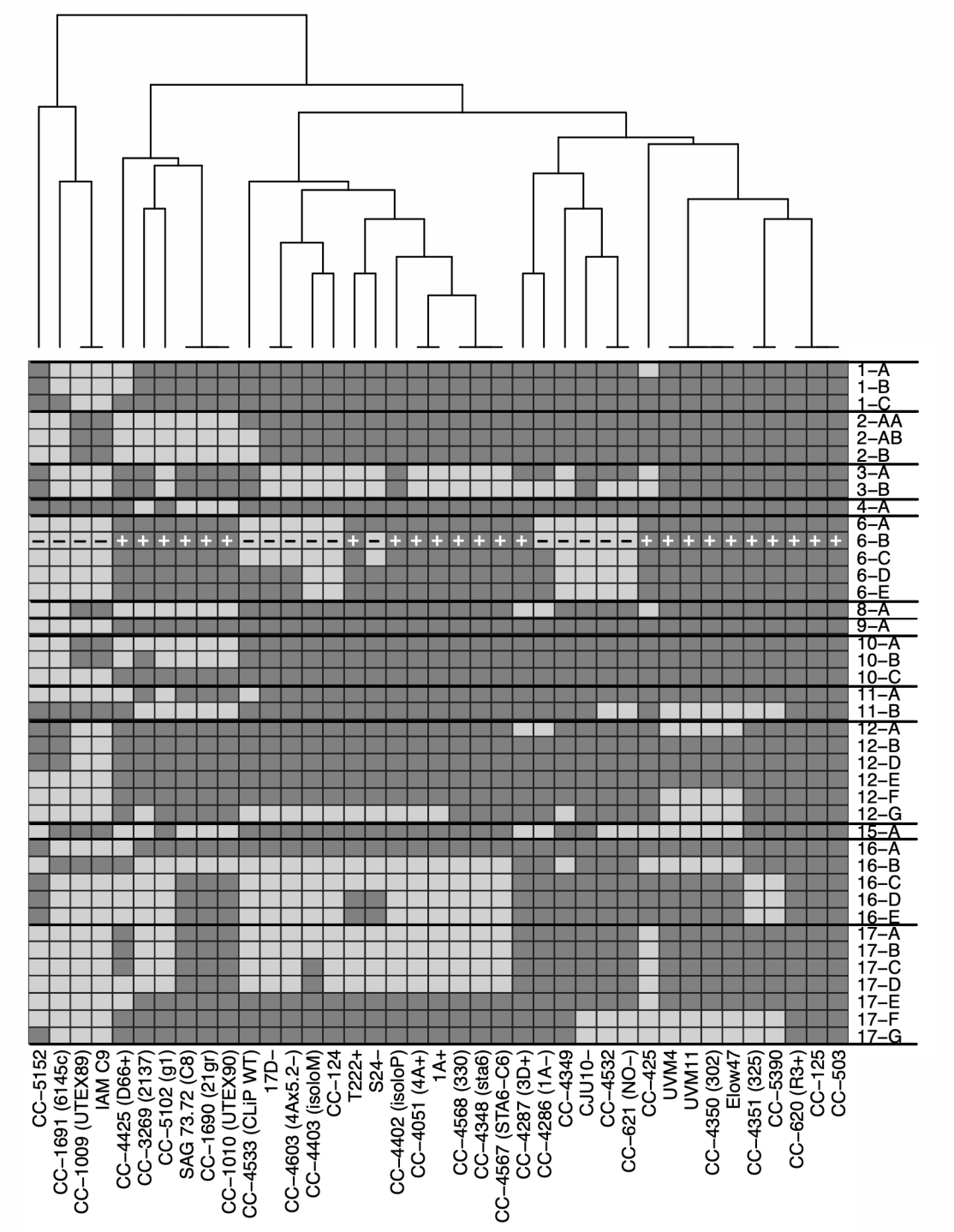

Figure 5.8 Distribution of two haplotypes in laboratory strains.
Each block represents a stretch of sequence that is either haplotype 1 (dark gray i.e. the same as CC-503) or 2 (light gray i.e. opposite of CC-503) in a given strain, with chromosomes shown on the y-axis. Strain mating types are shown by "+" and "–" symbols in block 6-B. For genomic coordinates of each block on the CC-4532 v6 genome, see Craig et al. (2022).

Although all ancestral and mutant variants segregating among laboratory strains are yet to be fully disentangled, studies of within- and between-haplotype diversity have been highly informative. More than 99% of single nucleotide polymorphisms (SNPs) called by Gallaher et al. (2015) were associated with between-haplotype variation, which as outlined are

expected to be ancestral differences. Slightly over a quarter of these variants were in coding sequence, more than half of which were synonymous, leaving ~65,000 variants that altered amino acids across haplotype 2 regions. Substantial gene expression and phenotypic differences were also reported between strains, and haplotype-specific regulatory and coding sequence variants are expected to underlie much of this variation. The remaining 1% of SNPs were found in within-haplotype comparisons across the 39 strains, corresponding to more than 4,000 single nucleotide mutations that presumably occurred in the laboratory. Approximately 200 of these were predicted to be loss-of-function mutations, including the famous *nit1-137* that abolishes nitrate reductase activity, found in the "137c" strains (CC-124, CC-125 and their descendants). Interestingly, the *nit2-137* nonsense mutation of the regulatory gene *NIT2*, found in CC-125, CC-503 and CC-4532, is not carried by CC-124 and several other 137c strains, which instead carry a TE insertion in exon 1.

Gallaher et al. (2015) also called 4,000 structural variants (i.e. variants >50 bp), with the most common class being deletions of on average ~5 kb, some of which were shown to affect genic sequence. Although some of the haplotype 2 regions have now been assembled in the CC-4532 v6 (Figure 5.1A) and CC-1690 assemblies, there has not yet been a systematic attempt to call structural variants between these regions and haplotype 1, and assessing between-haplotype structural differences will be an important next step in *Chlamydomonas* genome research. Some large differences are already obvious, such as in the rapidly evolving *NCL* gene family cluster (Boulouis et al. 2015). It lies partly within a region where CC-503 and CC-4532 carry alternative haplotypes on chromosome 15 (Figure 5.1B) and differs in gene copy number between the two assemblies (Craig et al. 2022).

Structural variants between copies of the same haplotype (i.e. structural mutations in the laboratory) were identified from the comparison of the CC-503 v6, CC-4532 v6 and CC-1690 assemblies (Section II). The ~100 large mutations (excluding TEs) specific to the CC-503 assembly are likely an unusual case, perhaps due to its defect in a genome maintenance pathway. More relevant to normal genome dynamics, ten structural mutations not caused by TEs partially or fully disrupt eight genes in CC-4532 v6. As expected, TE variants were more numerous, with 27 laboratory insertions identified in CC-503 v6, and 109 in CC-4532 v6. Remarkably, 86 TE insertions in CC-4532 v6 were of the same 15.4 kb *Gypsy* LTR element (*Gypsy-7a_cRei*), contributing ~1.3 Mb of unique sequence to the strain (i.e. an ~1% increase in genome size). Although only ten of the insertions were predicted to disrupt coding sequence, this demonstrates that substantial TE proliferation can occur in the laboratory. Utilizing Illumina re-sequencing data, similar proliferation of the *Gypsy* element was inferred in other laboratory strains, while others (including CC-503) showed no evidence of new insertions. Overlaying these results with known strain history, it appears that this specific TE family was ancestrally silenced but has become active independently in several lineages (Craig et al. 2022). Considering both v6 assemblies, the nonautonomous *Penelope*-like element *MRC1* was highly active in both strains and may generally be one of the most active TEs in the laboratory (Neupert et al. 2020), while several other families were found to be active at lower rates (Table 5.5).

Other unique structural variants that have arisen during laboratory culture were also identified by Flowers et al. (2015) using short-read coverage. This includes a duplication of several hundred kb on chromosome 13 in CC-1690, longer than the Nanopore read-lengths and hence collapsed in the CC-1690 assembly. Shorter duplications were also seen in CC-1010 (UTEX 90) and CC-407 (C-8), while another ~400 kb duplication on chromosome 1 was unique to CC-407. Large duplications were also among the mutations observed in the CC-4532 v6 assembly, and copy-number variants may be a substantial source of within-haplotype variation. Together with the variation present between the two alternative haplotypes, these strain-specific mutations and structural variants should reinforce a notion that has been variably accepted by the community: laboratory strains are neither isogenic, interchangeable, nor perfectly stable over time.

B. Population & species-level genomic variation

As of 2022, only 36 genetically distinct *Chlamydomonas* field isolates have been described, all of which have been subject to whole-genome re-sequencing (Jang and Ehrenreich 2012; Flowers et al. 2015; Craig et al. 2019). Apart from two sampled in Kyushu, Japan, all isolates were from eastern North America. The sampled isolates form three major geographic lineages, two in N. America (NA1 and NA2) and one in Japan (JPN). The first N. American group comprises the laboratory strains, CC-1373 (i.e. *Chlamydomonas smithii*, also from Massachusetts) and >20 isolates sampled from Quebec. The second comprises all other strains from Minnesota, Pennsylvania, North Carolina and Florida, although this group is very sparsely sampled and may in fact represent more than one genetic population. Each lineage is highly differentiated from one another, implying barriers to gene flow and the possibility of some degree of reproductive isolation. In approximate terms, pairwise diversity between two random isolates from within a lineage is ~2%, while pairwise diversity between two isolates from different lineages rises to slightly less than 3% (e.g. between the NA2 CC-1952/CC-2290 from Minnesota and the NA1 laboratory strains).

Flowers et al. (2015) explored the substantial variation present among field isolates with respect to potential loss-of-function polymorphisms. They found that 7.5% of genes harbored such polymorphisms in at least one isolate, although these were enriched in genes without *A. thaliana* homologs and it is possible that a non-negligible proportion were low quality models in the v5 annotation (Section III). Based on *de novo* assembly of reads that did not map to the reference genome, they also attempted to identify genes specific to a subset of strains and estimated that on average field isolate genomes contained 32 specific genes encoding proteins with recognizable domains, and many more without domains. A fraction may be associated with giant endogenous viral elements that are integrated to the genomes of a small subset of the field isolates (Moniruzzaman and Aylward 2021). In their analysis of coverage, Flowers et al. (2015) also found several large copy-number variants in field isolates, including a putative >400 kb duplication on chromosome 8 in CC-1952 and CC-1373. Finally, a PacBio assembly of CC-2931 (North Carolina) has been produced, which appears to carry two reciprocal translocations involving chromosomes 1, 6 and 10 (Craig 2021). They are probably unique to CC-2931 (possibly having occurred during laboratory culture) since

they are flanked by TEs that are highly active in this strain and are known to mediate rearrangements (López-Cortegano et al. 2022).

The above results show that, at the population and species levels, *Chlamydomonas* harbors a substantial functional diversity, possibly more than any other model organism. The community is now poised to fully exploit this diversity to obtain deeper knowledge on genome dynamics and on the function of genes. As we write these pages, genome assemblies of several field isolates are being analyzed in addition to the highly divergent CC-2931 mentioned above. Furthermore, a *Chlamydomonas* pan-genome initiative, funded as a JGI Community Sequencing Project, was initiated in 2021 to generate and compare high quality assemblies of a large number of divergent field isolates. This project targets isolates from a geographic range far broader than eastern North America, incorporating the Japanese *Chlamydomonas* (Nakada et al. 2014) and putative new isolates from California and France. With more fully-assembled genomes, correlation between genomic and phenotypic variation will be facilitated, for example in interbred populations generated by crossing widely different isolates.

C. Comparative genomics in the *Reinhardtinia* clade

Most of the comparative analyses performed between *C. reinhardtii* and closely related species have been performed at the protein level, which are discussed in Vol. 1 Chapter 4. These analyses may involve *V. carteri* and its multicellular relatives, which together with *C. reinhardtii* and its unicellular relatives belong to the core-*Reinhardtinia* clade (Nakada et al. 2016) (Vol. 1 Chapter 1). But *V. carteri* is separated from *C. reinhardtii* by ~230 million years of evolution (Herron et al. 2009), too much to apply the nucleotide-level approaches that can be powerful tools for refining structural annotations and detecting functional noncoding sequences. Until recently, the most closely related assembly was that of *C. sphaeroides* (Hirashima et al. 2016) but it was too fragmented for a refined analysis. In order to bridge this gap, Craig et al. (2021a) generated and annotated highly-contiguous assemblies for the two closest known relatives of *C. reinhardtii*, *C. incerta* and *C. schloesseri*, and one more distantly related unicellular species, *Edaphochlamys debaryana*. These assemblies were ~20-30 Mb larger than the *Chlamydomona*s genome, which could generally be attributed to higher repeat contents. Synteny was high and gene contents very similar between the three *Chlamydomonas* genomes, while the more gene-rich *E. debaryana* assembly showed a lesser synteny. Many other characteristic features of the *Chlamydomonas* genome were shared amongst all species, including high GC contents, long and abundant introns, highly diverse TE repertoires, and potentially a common centromeric organization (Section IV). As noted in Section III, comparison with these genomes led to the discovery of genes initially missed by the v5 annotations, highlighting the potential for interspecific comparative analyses to reveal biological novelty. They were also used to study the evolution of subtelomeric regions (Chaux-Jukic et al. 2021). Although Craig et al. (2021a) attempted to identify evolutionarily conserved noncoding elements using these resources, the identified sequences were much longer than the regulatory elements they hoped to capture. Efforts should thus be made to

isolate additional *Chlamydomonas* relatives to increase detection power, some of which could come out of field research undertaken as part of the pan-genome initiative.

VI. Online resources for the *Chlamydomonas* genome

Numerous web-based resources are available for *Chlamydomonas*, but their very nature is to evolve rapidly, not to mention the possibility that they disappear overnight due to funding cuts. The following description, valid as of February 2022, is thus neither complete nor final.

Just like the *Chlamydomonas* Resource Center is the major hub for strains, plasmids and methods, Phytozome (https://phytozome-next.jgi.doe.gov) is the main resource to access the most recent genomic data. Phytozome is by nature a comparative database, covering mostly land plants but also several green algae (nine Chlorophytes as of Feb 2022). As of v6, the organelle genomes are also included for *Chlamydomonas*. Every gene page gives access not only to the basic sequence information and alignment to homologs, but also to the functional annotation generated by JGI's multi-faceted automatic pipeline and by experts. The latter can include a gene symbol, deprecated aliases, a defline, comments and literature references. Also accessible are RNA-seq-based expression analyses (including co-expressed genes), prediction of intracellular targeting generated by TargetP and Predalgo (Emanuelsson et al. 2000; Tardif et al. 2012) and of transmembrane helices based on TMHMM (Krogh et al. 2001), as well as presence in various experimental datasets (interactome, experimental localization, plastid-ribosome pulldown, transcription factors and flagellar proteome databases, GEnome-scale Metabolic Modeling etc.), all based on previous genome versions. A link is provided to the CLiP library of insertional mutants (https://www.chlamylibrary.org) (Vol. 1 Chapter 17) and a description of the phenotypes. In Phytozome, proteins are grouped within families which can be used to retrieve multiple sequence alignments, trees and synteny data. In addition, the Phytozome genome browser presents as feature tracks a vast array of datasets allowing locus-level analyses, including sequence gaps, repeats, non-coding RNAgenes, alignment to other proteomes, gene models from previous annotations, sequence variants, CLiP mutation sites, alignment of EST and RNA-seq data, ribosome footprinting results, expression analysis etc. As in all JBrowse implementations, the arrowhead at the right of the track name allows changing of its configuration and downloading the local or chromosome-level track data.

As the community transitions from one version to another, older assemblies become obsolete, but it is still important to be able to retrieve the archived data for exploiting earlier work. Mycocosm hosts the version 2, 3 and 4 assemblies, accessible only via their direct url: https://mycocosm.jgi.doe.gov/Chlre2, Chlre3 or Chlre4. The main portal for JGI's algal genomes is now Phycocosm (https://phycocosm.jgi.doe.gov) which hosts hundreds of species including many Chlorophytes, such as *C. incerta* and *C. schloesseri*. Phycocosm is also comparative, and displays MCL protein clusters, a synteny browser and access to KBase with species trees. It also displays the organelle genomes and will in the future host data from the pan-genome initiative. The chloroplast genome of the Stern lab (archival) can also be found at the Resource Center http://www.chlamy.org/chloro. Another archival browser displays

transcriptome data (http://genomes.mcdb.ucla.edu/cgi-bin/hgTracks?db=chlRei4). *Chlamydomonas* genome data is also hosted at ENSEMBL (https://plants.ensembl.org/Chlamydomonas_reinhardtii for v5.5), KEGG (https://www.genome.jp/kegg-bin/show_organism?org=cre), Uniprot (https://www.uniprot.org/proteomes/UP000006906, which proteomics facilities often prefer to use as it is most compatible with their software) and the Plant Genome Database (https://www.plantgdb.org/CrGDB/), among other generalist databases. *Chlamydomonas* tRNAs are described in the PlantRNA database (http://seve.ibmp.unistra.fr/plantrna/), Transcription factors can be accessed via the Plant Transcription Factor Database (http://planttfdb.gao-lab.org/index.php?sp=Cre), and the flagellar proteome via http://chlamyfp.org/. BioCyc (https://biocyc.org/CHLAMY) and Mapman (https://mapman.gabipd.org/) display biological pathways. We anticipate that AlphaFold structural predictions will soon be available for all *Chlamydomonas* proteins, for example via Uniprot. Careful users of genome data will find that these resources usefully complete the view provided by Phytozome, and will uncover others using dedicated web searches. But remember the note of caution stressed in Latin by Blaby-Haas and Merchant (2019): whenever using genome data, "caveat emptor".

VII. Future Perspectives

At the time we are writing (05/2022), *Chlamydomonas* genomics is entering an exciting new era. With the completion of the v6 project, the reference assembly is expected to be near-complete with respect to genic sequence, and while an entirely gapless and highly accurate telomere-to-telomere assembly has yet to be published, this is on the horizon. Structural annotations are also reaching the highest quality standards, and while there will always remain scope for improvement, the increasing depth and variety of omics data being generated promises many new opportunities in this area (e.g. thoroughly annotating lncRNA genes). Importantly, the CC-4532 v6.1 annotation has formed the groundwork for assigning gene orthology and consistent nomenclature to the expanding number of genomes and annotations that are expected for new strains and closely related species in the coming years.

Genome evolution in *Chlamydomonas* is rapid and follows many routes, and we have outlined above the substantial variation among *Chlamydomonas* laboratory strains and field isolates, ranging from the level of single nucleotides to large duplications or rearrangements affecting entire chromosomes. While the reference genome will remain central to *Chlamydomonas* research, we expect many of the most interesting developments to come from the analysis of intraspecific diversity, in particular through the pan-genome initiative. Just like the availability of highly efficient genetic tools has spurred the advent of *Chlamydomonas* as a premier model organism in many fields of research, it is expected that further developments will largely rest on the improvement of resources for (multi-)genome level analyses.

Acknowledgments

RC's contributions were supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory (to Sabeeha Merchant) under U.S. Department of Energy Contract No. DE-AC02-05CH11231, and the Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/M010996/1]. OV's contributions were supported by the CNRS (UMR7141) and by the 'Initiative d'Excellence' program from the French State ['DYNAMO', ANR-11-LABX-0011-01].

REFERENCES

Antal M, Mougin A, Kis M, Boros E, Steger G, Jakab G, Solymosy F, Branlant C. 2000. Molecular characterization at the RNA and gene levels of U3 snoRNA from a unicellular green alga, *Chlamydomonas reinhardtii*. *Nucleic acids research* **28**: 2959-2968.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Aravind L, Balasubramanian S, Rao A. 2019. Unusual activity of a *Chlamydomonas* TET/JBP family enzyme. *Biochemistry* **58**: 3627-3629.

Asamizu E, Miura K, Kucho K, Inoue Y, Fukuzawa H, Ohyama K, Nakamura Y, Tabata S. 2000. Generation of expressed sequence tags from low-CO2 and high-CO2 adapted cells of *Chlamydomonas reinhardtii*. *DNA Res* **7**: 305-307.

Asamizu E, Nakamura Y, Sato S, Fukuzawa H, Tabata S. 1999. A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res* **6**: 369-373.

Babinger P, Volkl R, Cakstina I, Maftei A, Schmitt R. 2007. Maintenance DNA methyltransferase (Met1) and silencing of CpG-methylated foreign DNA in *Volvox carteri*. *Plant Mol Biol* **63**: 325-336.

Baier T, Jacobebbinghaus N, Einhaus A, Lauersen KJ, Kruse O. 2020. Introns mediate post-transcriptional enhancement of nuclear gene expression in the green microalga *Chlamydomonas reinhardtii*. *PLoS Genet* **16**: e1008944.

Bao W, Jurka J. 2013. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob DNA* **4**: 12.

Barahimipour R, Strenkert D, Neupert J, Schroda M, Merchant SS, Bock R. 2015. Dissecting the contributions of GC content and codon usage to gene expression in the model alga *Chlamydomonas reinhardtii*. *Plant J* **84**: 704-717.

Bell SA, Shen C, Brown A, Hunt AG. 2016. Experimental genome-wide determination of RNA polyadenylation in *Chlamydomonas reinhardtii*. *PLoS One* **11**: e0146107.

Blaby IK, Blaby-Haas CE. 2017. Genomics and functional genomics in *Chlamydomonas reinhardtii*. In *Chlamydomonas: Molecular genetics and physiology*, (ed. M Hippler). Springer.

Blaby IK, Blaby-Haas CE, Tourasse N, Hom EF, Lopez D, Aksoy M, Grossman A, Umen J, Dutcher S, Porter M et al. 2014. The *Chlamydomonas* genome project: a decade on. *Trends in Plant Science* **19**: 672-680.

Blaby-Haas CE, Merchant SS. 2019. Comparative and functional algal genomics. *Annual Review of Plant Biology* **70**: 605-638.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S et al. 2012. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol* **13**.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* **25**: 106-110.

Böhne A, Zhou Q, Darras A, Schmidt C, Schartl M, Galiana-Arnoux D, Volff JN. 2012. Zisupton—a novel superfamily of DNA transposable elements recently active in fish. *Mol Biol Evol* **29**: 631-645.

Boulouis A, Drapier D, Razafimanantsoa H, Wostrikoff K, Tourasse NJ, Pascal K, Girard-Bascou J, Vallon O, Wollman FA, Choquet Y. 2015. Spontaneous dominant mutations in *Chlamydomonas* highlight ongoing evolution by gene diversification. *Plant Cell* **27**: 984-1001.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897-907.

Casas-Mollano JA, Rohr J, Kim EJ, Balassa E, van Dijk K, Cerutti H. 2008. Diversification of the core RNA interference machinery in *Chlamydomonas reinhardtii* and the role of DCL1 in transposon silencing. *Genetics* **179**: 69-81.

Cavaiuolo M, Kuras R, Wollman FA, Choquet Y, Vallon O. 2017. Small RNA profiling in *Chlamydomonas*: insights into chloroplast RNA metabolism. *Nucleic Acids Res* **45**: 10783-10799.

Cenkci B, Petersen JL, Small GD. 2003. REX1, a novel gene required for DNA repair. *J Biol Chem* **278**: 22574-22577.

Chang CH, Chavan A, Palladino J, Wei XL, Martins NMC, Santinello B, Chen CC, Erceg J, Beliveau BJ, Wu CT et al. 2019. Islands of retroelements are major components of *Drosophila* centromeres. *PLoS Biol* **17**.

Chaux-Jukic F, O'Donnell S, Craig RJ, Eberhard S, Vallon O, Xu Z. 2021. Architecture and evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*. *Nucleic Acids Res* doi:10.1093/nar/gkab534.

Chen CL, Chen CJ, Vallon O, Huang ZP, Zhou H, Qu LH. 2008. Genomewide analysis of box C/D and box H/ACA snoRNAs in *Chlamydomonas reinhardtii* reveals an extensive organization into intronic gene clusters. *Genetics* **179**: 21-30.

Cognat V, Deragon JM, Vinogradova E, Salinas T, Remacle C, Marechal-Drouard L. 2008. On the evolution and expression of *Chlamydomonas reinhardtii* nucleus-encoded transfer RNA genes. *Genetics* **179**: 113-123.

Craig RJ. 2021. The evolutionary genomics of *Chlamydomonas*. PhD thesis, University of Edinburgh.

Craig RJ, Böndel KB, Arakawa K, Nakada T, Ito T, Bell G, Colegrave N, Keightley PD, Ness RW. 2019. Patterns of population structure and complex haplotype sharing among field isolates of the green alga *Chlamydomonas reinhardtii*. *Mol Ecol* **28**: 3977-3993.

Craig RJ, Gallaher SD, Shu S, Salome PA, Jenkins JW, Blaby-Haas CE, Purvine SO, O'Donnell S, Barry K, Grimwood J et al. 2022. The Chlamydomonas Genome Project, Version 6: reference assemblies for mating type *plus* and *minus* strains reveal extensive structural mutation in the laboratory *Biorxiv*.

Craig RJ, Hasan AR, Ness RW, Keightley PD. 2021a. Comparative genomics of *Chlamydomonas*. *Plant Cell* **33**: 1016-1041.

Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. 2021b. An ancient clade of *Penelope*-like retroelements with permuted domains is present in the green lineage and protists, and dominates many invertebrate genomes. *Mol Biol Evol* **38**: 5005-5020.

Croft MT, Moulin M, Webb ME, Smith AG. 2007. Thiamine biosynthesis in algae is regulated by riboswitches. *Proc Natl Acad Sci U S A* **104**: 20770-20775.

Cross FR. 2015. Tying down loose ends in the *Chlamydomonas* genome: functional significance of abundant upstream open reading frames. *G3 (Bethesda)* **6**: 435-446.

Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* **7**.

Darlix JL, Rochaix JD. 1981. Nucleotide sequence and structure of cytoplasmic 5S RNA and 5.8S RNA of *Chlamydomonas reinhardii*. *Nucleic Acids Res* **9**: 1291-1299.

Davies DR. 1972. Cell wall organisation in *Chlamydomonas reinhardi*. The role of extra-nuclear systems. *Mol Gen Genet* **115**: 334-348.

Day A. 1995. A transposon-like sequence with short terminal inverted repeats in the nuclear genome of *Chlamydomonas reinhardtii*. *Plant Mol Biol* **28**: 437-442.

Day A, Rochaix JD. 1991. A transposon with an unusual LTR arrangement from *Chlamydomonas reinhardtii* contains an internal tandem array of 76 bp repeats. *Nucleic Acids Res* **19**: 1259-1266.

Day A, Schirmerrahire M, Kuchka MR, Mayfield SP, Rochaix JD. 1988. A transposon with an unusual arrangement of long terminal repeats in the green alga *Chlamydomonas reinhardtii*. *EMBO Journal* **7**: 1917-1927.

De Hoff PL, Ferris P, Olson BJSC, Miyagi A, Geng S, Umen JG. 2013. Species and population level molecular profiling reveals cryptic recombination and emergent asymmetry in the dimorphic mating locus of *C. reinhardtii*. *PLoS Genet* **9**.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285-311.

Eberhard S, Valuchova S, Ravat J, Fulnecek J, Jolivet P, Bujaldon S, Lemaire SD, Wollman FA, Teixeira MT, Riha K et al. 2019. Molecular characterization of *Chlamydomonas reinhardtii* telomeres and telomerase mutants. *Life Sci Alliance* **2**.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**: 1005-1016.

Falciatore A, Merendino L, Barneche F, Ceol M, Meskauskiene R, Apel K, Rochaix JD. 2005. The FLP proteins act as regulators of chlorophyll synthesis in response to light and plastid signals in *Chlamydomonas*. *Genes Dev* **19**: 176-187.

Farlow A, Meduri E, Schlotterer C. 2011. DNA double-strand break repair and the evolution of intron density. *Trends Genet* **27**: 1-6.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME et al. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**: 8689-8694.

Ferenczi A, Chew YP, Kroll E, von Koppenfels C, Hudson A, Molnar A. 2021. Mechanistic and genetic basis of single-strand templated repair at Cas12a-induced DNA breaks in *Chlamydomonas reinhardtii*. *Nat Commun* **12**: 6751.

Ferris P, Olson BJ, De Hoff PL, Douglass S, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, Schmutz J et al. 2010. Evolution of an expanded sex-determining locus in *Volvox*. *Science* **328**: 351-354.

Ferris PJ. 1989. Characterization of a *Chlamydomonas* transposon, *Gulliver*, resembling those in higher-plants. *Genetics* **122**: 363-377.

Ferris PJ, Goodenough UW. 1997. Mating type in *Chlamydomonas* is specified by *mid*, the minus-dominance gene. *Genetics* **146**: 859-869.

Ferris PJ, Woessner JP, Goodenough UW. 1996. A sex recognition glycoprotein is encoded by the *plus* mating-type gene *fus1* of *Chlamydomonas reinhardtii*. *Mol Biol Cell* **7**: 1235-1248.

Fischer N, Rochaix JD. 2001. The flanking regions of *PsaD* drive efficient gene expression in the nucleus of the green alga *Chlamydomonas reinhardtii*. *Mol Genet Genomics* **265**: 888-894.

Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH et al. 2015. Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. *Plant Cell* **27**: 2353-2369.

Fu Y, Luo GZ, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Dore LC et al. 2015. N^6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell* **161**: 879-892.

Fukuzawa H, Miura K, Ishizaki K, Kucho KI, Saito T, Kohinata T, Ohyama K. 2001. *Ccm1*, a regulatory gene controlling the induction of a carbon-concentrating mechanism in *Chlamydomonas reinhardtii* by sensing CO2 availability. *Proc Natl Acad Sci U S A* **98**: 5347-5352.

Fulnečková J, Hasíková T, Fajkus J, Lukešová A, Eliáš M, Sýkorová E. 2012. Dynamic evolution of telomeric sequences in the green algal order Chlamydomonadales. *Genome Biol Evo* **4**: 248-264.

Gallaher SD, Craig RJ, Ganesan I, Purvine SO, McCorkle S, Grimwood J, Strenkert D, Davidi L, Roth MS, Jeffers TL et al. 2021. Widespread polycistronic gene expression in green algae. *Proc Natl Acad Sci U S A* **118**: e2017714118.

Gallaher SD, Fitz-Gibbon ST, Glaesener AG, Pellegrini M, Merchant SS. 2015. *Chlamydomonas* genome resource for laboratory strains reveals a mosaic of sequence variation, identifies true strain histories, and enables strain-specific studies. *Plant Cell* **27**: 2335-2352.

Gallaher SD, Fitz-Gibbon ST, Strenkert D, Purvine SO, Pellegrini M, Merchant SS. 2018. High-throughput sequencing of the chloroplast and mitochondrion of *Chlamydomonas reinhardtii* to generate improved *de novo* assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. *Plant J* **93**: 545-565.

Glöckner G, Heidel AJ. 2009. Centromere sequence and dynamics in *Dictyostelium discoideum*. *Nucleic Acids Res* **37**: 1809-1816.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92-100.

Goodwin TJ, Poulter RT. 2001. The DIRS1 group of retrotransposons. *Mol Biol Evol* **18**: 2067-2082.

Goodwin TJ, Poulter RT. 2004. A new group of tyrosine recombinase-encoding retrotransposons. *Mol Biol Evol* **21**: 746-759.

Graham JE, Spanier JG, Jarvik JW. 1995. Isolation and characterization of *Pioneer1*, a novel *Chlamydomonas* transposable element. *Current Genetics* **28**: 429-436.

Grossman AR, Harris EE, Hauser C, Lefebvre PA, Martinez D, Rokhsar D, Shrager J, Silflow CD, Stern D, Vallon O et al. 2003. *Chlamydomonas reinhardtii* at the crossroads of genomics. *Eukaryotic cell* **2**: 1137-1150.

Haeusler RA, Engelke DR. 2006. Spatial organization of transcription by RNA polymerase III. *Nucleic Acids Res* **34**: 4826-4836.

Hamaji T, Kawai-Toyooka H, Uchimura H, Suzuki M, Noguchi H, Minakuchi Y, Toyoda A, Fujiyama A, Miyagishima S, Umen JG et al. 2018. Anisogamy evolved with a reduced sex-determining region in volvocine green algae. *Communications Biology* **1**.

Harris EH. 2009. *The Chlamydomonas Sourcebook (Second Edition): Introduction to Chlamydomonas and Its laboratory use*. Academic Press.

Hasan AR, Duggal JK, Ness RW. 2019. Consequences of recombination for the evolution of the mating type locus in *Chlamydomonas reinhardtii*. *New Phytologist* **224**: 1339-1348.

Hasan AR, Ness RW. 2020. Recombination rate variation and infrequent sex influence genetic diversity in *Chlamydomonas reinhardtii*. *Genome Biol Evol* **12**: 370-380.
Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. *Proc Natl Acad Sci U S A* **106**: 3254-3258.
Higashiyama T, Noutoshi Y, Fujie M, Yamada T. 1997. Zepp, a LINE-like retrotransposon accumulated in the *Chlorella* telomeric region. *EMBO J* **16**: 3715-3723.
Hirashima T, Tajima N, Sato N. 2016. Draft genome sequences of four species of *Chlamydomonas* containing phosphatidylcholine. *Microbiol Resour Ann* **4**.
Howell SH. 1972. The differential synthesis and degradation of ribosomnal DNA during the vegetative cell-cycle in *Chlamydomonas reinhardi*. *Nature New Biology* **240**: 264-267.
Hyams J, Davies DR. 1972. Induction and characterization of cell-wall mutants of *Chlamydomonas reinhardi*. *Mutat Res* **14**: 381-&.
Iyer LM, Zhang DP, de Souza RF, Pukkila PJ, Rao A, Aravind L. 2014. Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. *Proc Natl Acad Sci U S A* **111**: 1676-1683.
Jain M, Shrager J, Harris EH, Halbrook R, Grossman AR, Hauser C, Vallon O. 2007. EST assembly supported by a draft genome sequence: an analysis of the *Chlamydomonas reinhardtii* transcriptome. *Nucleic Acids Res* **35**: 2074-2083.
Jakab G, Kis M, Solymosy F. 1992. Nucleotide sequence of U5 RNA from a green alga, *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **20**: 5224.
Jakab G, Mougin A, Kis M, Pollak T, Antal M, Branlant C, Solymosy F. 1997. *Chlamydomonas* U2, U4 and U6 snRNAs. An evolutionary conserved putative third interaction between U4 and U6 snRNAs which has a counterpart in the $U4_{atac}$-$U6_{atac}$ snRNA duplex. *Biochimie* **79**: 387-395.
Jang H, Ehrenreich IM. 2012. Genome-wide characterization of genetic variation in the unicellular, green alga *Chlamydomonas reinhardtii*. *PLoS One* **7**: e41307.
Kang Y, Mitchell DR. 1998. An intronic enhancer is required for deflagellation-induced transcriptional regulation of a *Chlamydomonas reinhardtii* dynein gene. *Mol Biol Cell* **9**: 3085-3094.
Kapitonov VV, Jurka J. 2004. L1-1_CR, a family of L1-like non-LTR retrotransposons from the green algae genome. *Repbase Reports* **4**: 39.
Kapitonov VV, Jurka J. 2006. Gulliver, a family of autonomous hAT transposons from the green algae genome. *Repbase Reports* **6**: 227.
Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* **9**: 411-412; author reply 414.
Kathir P, LaVoie M, Brazelton WJ, Haas NA, Lefebvre PA, Silflow CD. 2003. Molecular map of the *Chlamydomonas reinhardtii* nuclear genome. *Eukaryot Cell* **2**: 362-379.
Keskiaho K, Hieta R, Sormunen R, Myllyharju J. 2007. *Chlamydomonas reinhardtii* has multiple prolyl 4-hydroxylases, one of which is essential for proper cell wall assembly. *Plant Cell* **19**: 256-269.
Kianianmomeni A, Ong CS, Ratsch G, Hallmann A. 2014. Genome-wide analysis of alternative splicing in *Volvox carteri*. *Bmc Genomics* **15**: 1117.
Kim KS, Kustu S, Inwood W. 2006. Natural history of transposition in the green alga *Chlamydomonas reinhardtii*: Use of the AMT4 locus as an experimental system. *Genetics* **173**: 2005-2019.
Kis M, Jakab G, Pollak T, Branlant C, Solymosy F. 1993. Nucleotide sequence of U1 RNA from a green alga, *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **21**: 2255.
Kojima KK, Fujiwara H. 2005. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15**: 1106-1117.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567-580.
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639-1645.
Kwan AL, Li L, Kulp DC, Dutcher SK, Stormo GD. 2009. Improving gene-finding in *Chlamydomonas reinhardtii*:GreenGenie2. *Bmc Genomics* **10**: 210.
Labadorf A, Link A, Rogers MF, Thomas J, Reddy AS, Ben-Hur A. 2010. Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. *Bmc Genomics* **11**: 114.
Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**: e1001388.
Li H, Wang Y, Chen M, Xiao P, Hu C, Zeng Z, Wang C, Wang J, Hu Z. 2016. Genome-wide long non-coding RNA screening, identification and characterization in a model microorganism *Chlamydomonas reinhardtii*. *Sci Rep* **6**: 34109.
Li JB, Lin S, Jia H, Wu H, Roe BA, Kulp D, Stormo GD, Dutcher SK. 2003. Analysis of *Chlamydomonas reinhardtii* genome structure using large-scale sequencing of regions on linkage groups I and III. *J Eukaryot Microbiol* **50**: 145-155.
Li XQ, Du D. 2014. Motif types, motif locations and base composition patterns around the RNA polyadenylation site in microorganisms, plants and animals. *BMC Evol Biol* **14**: 162.
Lin H, Cliften PF, Dutcher SK. 2018a. MAPINS, a highly efficient detection method that identifies insertional mutations and complex DNA rearrangements. *Plant Physiol* **178**: 1436-1447.
Lin H, Dutcher SK. 2015. Genetic and genomic approaches to identify genes involved in flagellar assembly in *Chlamydomonas reinhardtii*. *Methods Cell Biol* **127**: 349-386.
Lin H, Miller ML, Granas DM, Dutcher SK. 2013. Whole genome sequencing identifies a deletion in protein phosphatase 2A that affects its stability and localization in *Chlamydomonas reinhardtii*. *PLoS Genet* **9**: e1003841.
Lin H, Zhang Z, Iomini C, Dutcher SK. 2018b. Identifying RNA splicing factors using *IFT* genes in *Chlamydomonas reinhardtii*. *Open Biol* **8**.
Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nat Ecol Evol* **2**: 164-173.
Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. 2019. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat Commun* **10**: 2449.
Liu Y, Li S, Chen Y, Kimberlin AN, Cahoon EB, Yu B. 2016. snRNA 3' End Processing by a CPSF73-Containing Complex Essential for Development in Arabidopsis. *PLoS biology* **14**: e1002571.
Lodha M, Schulz-Raffelt M, Schroda M. 2008. A new assay for promoter analysis in *Chlamydomonas* reveals roles for heat shock elements and the TATA box in *HSP70A* promoter-mediated activation of transgene expression. *Eukaryot Cell* **7**: 172-176.
Lopez D, Hamaji T, Kropat J, De Hoff P, Morselli M, Rubbi L, Fitz-Gibbon S, Gallaher SD, Merchant SS, Umen J et al. 2015. Dynamic changes in the transcriptome and methylome of *Chlamydomonas reinhardtii* throughout Its life cycle. *Plant Physiol* **169**: 2730-2743.

López-Cortegano E, Craig RJ, Chebib J, Balogun EJ, Keightley PD. 2022. Rates and spectra of *de novo* structural mutation in *Chlamydomonas reinhardtii*. *Biorxiv*.
López-Cortegano E, Craig RJ, Chebib J, Samuels T, Morgan AD, Kraemer SA, Bondel KB, Ness RW, Colegrave N, Keightley PD. 2021. De novo mutation rate variation and its determinants in *Chlamydomonas*. *Mol Biol Evol* doi:10.1093/molbev/msab140.
Loppes R, Radoux M. 2002. Two short regions of the promoter are essential for activation and repression of the nitrate reductase gene in *Chlamydomonas reinhardtii*. *Mol Genet Genomics* **268**: 42-48.
Lu H, Davis AJ. 2021. Human RecQ helicases in DNA double-strand break repair. *Front Cell Dev Biol* **9**: 640755.
Merchant SS Prochnik SE Vallon O Harris EH Karpowicz SJ Witman GB Terry A Salamov A Fritz-Laylin LK Marechal-Drouard L et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**: 245-250.
Moniruzzaman M, Aylward FO. 2021. Endogenous giant viruses shape intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. *Biorxiv* doi:https://doi.org/10.1101/2021.11.30.470594.
Moseley JL, Page MD, Alder NP, Eriksson M, Quinn J, Soto F, Theg SM, Hippler M, Merchant S. 2002. Reciprocal expression of two candidate di-iron enzymes affecting photosystem I and light-harvesting complex accumulation. *Plant Cell* **14**: 673-688.
Müller SY, Matthews NE, Valli AA, Baulcombe DC. 2020. The small RNA locus map for *Chlamydomonas reinhardtii*. *PLoS One* **15**: e0242516.
Nakada T, Ito T, Tomita M. 2016. 18S ribosomal RNA gene phylogeny of a colonial volvocalean lineage (*Tetrabaenaceae-Goniaceae-Volvocaceae, Volvocales, Chlorophyceae*) and its close relatives. *The Journal of Japanese Botany* **91**: 345-354.
Nakada T, Tsuchida Y, Arakawa K, Ito T, Tomita M. 2014. Hybridization between Japanese and North American *Chlamydomonas reinhardtii* (Volvocales, Chlorophyceae). *Phycol Res* **62**: 232-236.
Naya H, Romero H, Carels N, Zavala A, Musto H. 2001. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. *FEBS Lett* **501**: 127-130.
Ness RW, Kraemer SA, Colegrave N, Keightley PD. 2016. Direct estimate of the spontaneous mutation rate uncovers the effects of drift and recombination in the *Chlamydomonas reinhardtii* plastid genome. *Mol Biol Evol* **33**: 800-808.
Ness RW, Morgan AD, Colegrave N, Keightley PD. 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics* **192**: 1447-1454.
Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Res* **25**: 1739-1749.
Neupert J, Gallaher SD, Lu Y, Strenkert D, Segal N, Barahimipour R, Fitz-Gibbon ST, Schroda M, Merchant SS, Bock R. 2020. An epigenetic gene silencing pathway selectively acting on transgenic DNA in the green alga *Chlamydomonas*. *Nat Commun* **11**: 6269.
Ngan CY, Wong CH, Choi C, Yoshinaga Y, Louie K, Jia J, Chen C, Bowen B, Cheng H, Leonelli L et al. 2015. Lineage-specific chromatin signatures reveal a regulator of lipid metabolism in microalgae. *Nat Plants* **1**: 15107.
Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV, Xiang Y, Weeks DP, Hatfield DL, Gladyshev VN. 2002. Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *The EMBO journal* **21**: 3681-3693.

O'Donnell S, Chaux F, Fischer G. 2020. Highly Contiguous Nanopore Genome Assembly of *Chlamydomonas reinhardtii* CC-1690. *Microbiology resource announcements* **9**.
Ozawa SI, Cavaiuolo M, Jarrige D, Kuras R, Rutgers M, Eberhard S, Drapier D, Wollman FA, Choquet Y. 2020. The OPR protein MTHI1 controls the expression of two different subunits of ATP synthase CFo in *Chlamydomonas reinhardtii*. *Plant Cell* **32**: 1179-1203.
Pandey M, Stormo GD, Dutcher SK. 2020. Alternative splicing during the *Chlamydomonas reinhardtii* cell cycle. *G3 (Bethesda)* **10**: 3797-3810.
Perez-Alegre M, Dubus A, Fernandez E. 2005. REM1, a new type of long terminal repeat retrotransposon in *Chlamydomonas reinhardtii*. *Mol Cell Biol* **25**: 10628-10638.
Petracek ME, Lefebvre PA, Silflow CD, Berman J. 1990. *Chlamydomonas* telomere sequences are A+T-rich but contain three consecutive G-C base pairs. *Proc Natl Acad Sci U S A* **87**: 8222-8226.
Philippsen GS, Avaca-Crusca JS, Araujo APU, DeMarco R. 2016. Distribution patterns and impact of transposable elements in genes of green algae. *Gene* **594**: 151-159.
Pinhal D, Yoshimura TS, Araki CS, Martins C. 2011. The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. *BMC Evol Biol* **11**: 151.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32-42.
Poulter RT, Goodwin TJ. 2005. DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet Genome Res* **110**: 575-588.
Preuss D, Mets L. 2002. Plant centromere functions defined by tetrad analysis and artificial chromosomes. *Plant Physiol* **129**: 421-422.
Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**: 223-226.
Proschold T, Harris EH, Coleman AW. 2005. Portrait of a species: *Chlamydomonas reinhardtii*. *Genetics* **170**: 1601-1610.
Quinn JM, Merchant S. 1995. Two copper-responsive elements associated with the Chlamydomonas Cyc6 gene function as targets for transcriptional activators. *Plant Cell* **7**: 623-628.
Raj-Kumar PK, Vallon O, Liang C. 2017. *In silico* analysis of the sequence features responsible for alternatively spliced introns in the model green alga *Chlamydomonas reinhardtii*. *Plant Mol Biol* **94**: 253-265.
Rao M, Carlson BA, Novoselov SV, Weeks DP, Gladyshev VN, Hatfield DL. 2003. *Chlamydomonas reinhardtii* selenocysteine tRNA[Ser]Sec. *RNA* **9**: 923-930.
Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**: 2279-2286.
Rymarquis LA, Handley JM, Thomas M, Stern DB. 2005. Beyond complementation. Map-based cloning in *Chlamydomonas reinhardtii*. *Plant Physiol* **137**: 557-566.
Salinas-Giegé T, Cavaiuolo M, Cognat V, Ubrig E, Remacle C, Duchene AM, Vallon O, Marechal-Drouard L. 2017. Polycytidylation of mitochondrial mRNAs in *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **45**: 12963-12973.
Salomé PA, Merchant SS. 2019. A Series of fortunate events: Introducing *Chlamydomonas* as a reference organism. *Plant Cell* **31**: 1682-1707.
Schnell RA, Lefebvre PA. 1993. Isolation of the *Chlamydomonas* regulatory gene NIT2 by transposon tagging. *Genetics* **134**: 737-747.

Schroda M, Vallon O, Whitelegge JP, Beck CF, Wollman FA. 2001. The chloroplastic GrpE homolog of Chlamydomonas: two isoforms generated by differential splicing. *Plant Cell* **13**: 2823-2839.
Shrager J, Hauser C, Chang CW, Harris EH, Davies J, McDermott J, Tamse R, Zhang Z, Grossman AR. 2003. *Chlamydomonas reinhardtii* Genome Project. A guide to the generation and use of the cDNA information. *Plant Physiol* **131**: 401-408.
Silflow CD, Chisholm RL, Conner TW, Ranum LP. 1985. The two alpha-tubulin genes of *Chlamydomonas reinhardi* code for slightly different proteins. *Mol Cell Biol* **5**: 2389-2398.
Sinturel F, Navickas A, Wery M, Descrimes M, Morillon A, Torchet C, Benard L. 2015. Cytoplasmic control of sense-antisense mRNA pairs. *Cell Rep* **12**: 1853-1864.
Specht M, Stanke M, Terashima M, Naumann-Busch B, Janssen I, Hohner R, Hom EF, Liang C, Hippler M. 2011. Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the *Chlamydomonas reinhardtii* genome. *Proteomics* **11**: 1814-1823.
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.
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 Bioinformatics* **7**: 62.
Strenkert D, Yildirim A, Yan J, Yoshinaga Y, Pellegrini M, O''Malley C, Merchant SS, Umen J, G. . 2021. Genome wide profiling of histone H3 lysine 4 methylation during the Chlamydomonas cell cycle reveals stable and dynamic properties of lysine 4 trimethylation at gene promoters and near ubiquitous lysine 4 monomethylation. *Biorxiv* doi: https://doi.org/10.1101/2021.09.19.460975.
Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A* **109**: 18488-18492.
Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugiere S, Hippler M, Ferro M, Bruley C, Peltier G et al. 2012. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol Biol Evol* **29**: 3625-3639.
Tulin F, Cross FR. 2016. Patching holes in the *Chlamydomonas* genome. *G3 (Bethesda)* **6**: 1899-1910.
Valli AA, Santos BA, Hnatova S, Bassett AR, Molnar A, Chung BY, Baulcombe DC. 2016. Most microRNAs in the single-cell alga *Chlamydomonas reinhardtii* are produced by Dicer-like 3-mediated cleavage of introns and untranslated regions of coding RNAs. *Genome Res* **26**: 519-529.
van Dijk K, Xu H, Cerutti H. 2006. Epigenetic Silencing of transposons in the green alga *Chlamydomonas reinhardtii*. In *Small RNAs: Analysis and Regulatory Functions*, doi:10.1007/978-3-540-28130-6_8 (ed. W Nellen, C Hammann), pp. 159-178. Springer Berlin Heidelberg, Berlin, Heidelberg.
Wang SC, Schnell RA, Lefebvre PA. 1998. Isolation and characterization of a new transposable element in *Chlamydomonas reinhardtii*. *Plant Molecular Biology* **38**: 681-687.
Weissman JL, Fagan WF, Johnson PLF. 2019. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genet* **15**: e1008493.
Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. *Annu Rev Genet* **54**: 539-561.
Woessner JP, Goodenough UW. 1994. Volvocine cell walls and their constituent glycoproteins: an evolutionary perspective. *Protoplasma* **181**: 245-258.

Wu-Scharf D, Jeong B, Zhang C, Cerutti H. 2000. Transgene and transposon silencing in *Chlamydomonas reinhardtii* by a DEAH-box RNA helicase. *Science* **290**: 1159-1162.

Xue JH, Chen GD, Hao F, Chen H, Fang Z, Chen FF, Pang B, Yang QL, Wei X, Fan QQ et al. 2019. A vitamin-C-derived DNA modification catalysed by an algal TET homologue. *Nature* **569**: 581-585.

Yamamoto K, Hamaji T, Kawai-Toyooka H, Matsuzaki R, Takahashi F, Nishimura Y, Kawachi M, Noguchi H, Minakuchi Y, Umen JG et al. 2021. Three genomes in the algal genus *Volvox* reveal the fate of a haploid sex-determining region after a transition to homothallism. *Proc Natl Acad Sci U S A* **118**.

Yuan YW, Wessler SR. 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci U S A* **108**: 7884-7889.

Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang XJ, Qi Y. 2007. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev* **21**: 1190-1203.

Zhao Z, Guo C, Sutharzan S, Li P, Echt CS, Zhang J, Liang C. 2014a. Genome-wide analysis of tandem repeats in plants and green algae. *G3 (Bethesda)* **4**: 67-78.

Zhao Z, Wu X, Kumar PK, Dong M, Ji G, Li QQ, Liang C. 2014b. Bioinformatics analysis of alternative polyadenylation in green alga *Chlamydomonas reinhardtii* using transcriptome sequences from three different sequencing platforms. *G3 (Bethesda)* **4**: 871-883.

Zhu S, Ye W, Ye L, Fu H, Ye C, Xiao X, Ji Y, Lin W, Ji G, Wu X. 2020. PlantAPAdb: a comprehensive database for alternative polyadenylation sites in plants. *Plant Physiol* **182**: 228-242.

Zorin B, Hegemann P, Sizova I. 2005. Nuclear-gene targeting by using single-stranded DNA avoids illegitimate DNA integration in *Chlamydomonas reinhardtii*. *Eukaryot Cell* **4**: 1264-1272.