The Evolutionary Genomics of *Chlamydomonas*

THE UNIVERSITY
of EDINBURGH

Rory J. Craig

Submitted for the degree of Doctor of Philosophy
Institute of Evolutionary Biology
University of Edinburgh
2021

Declaration

I declare that this thesis was composed by myself and that this work has not been submitted for any other degree or professional qualification. The work presented in the thesis (including work that has been published elsewhere) is my own, expect where explicitly stated in the preface of each results chapter (Chapters 2-5).

Acknowledgements

I have been lucky enough to spend considerable parts of the past four or so years in four different countries and three universities, and the number of people who have helped, influenced and supported me is substantial. I will try my best to thank you all either by name or at least association. I first thank Peter Keightley for his guidance and friendship, it's been a pleasure to have you as a supervisor. Your mentorship has made my PhD a thoroughly rewarding experience, and I'm especially grateful that you've been so accommodating as I've followed various tangents and strayed increasingly from population genetics. I thank Nick Colegrave for all of his help over the years and for introducing me to chlamy in the first place (even if I wasn't all that keen on it at the time!). I thank Rob Ness, my unofficial third supervisor, whose guidance and excellent ideas have been invaluable throughout. I also sincerely thank my funders, the EASTBIO Doctoral Training Partnership/BBSRC, without whose financial and administrative support this work would not have been possible. I thank David Finnegan and Stephen Wright for their time in examining an earlier version of this thesis and for their thoughtful comments and suggestions which substantially improved this work.

Going further back, I thank Paul Rogers, who played a large role in my initial interest in biology. I thank Paul Sharp for his excellent supervision and teaching (also Peter and Darren Obbard) on the Honours evolutionary biology course. I thank Hans Ellegren for employing me on the back of cold email, I learnt an enormous amount about how to conduct research in my time in your lab. I thank the many people who made my year in Uppsala such a rewarding experience, Ludo, Jelmer, Venkat, Carina, Paulina, Homa, Linnéa, and several others. I especially thank Alex Suh for his mentorship and friendship, and everyone in the TE jamboree group for their virtual company over the past year.

I thank everyone from the Evolutionary Genetics group at Edinburgh, especially Deborah and Brian who have been such a wonderful source of inspiration (as well as interesting anecdotes!). I thank Konrad, Susie, Matty and many others. I thank the members of the Keightley lab past and present, Katharina, Susanne, Tom, Ben, Jobran and Eugenio, for their friendship and everything that they've taught me (including how to function in a laboratory). I thank the many friends I have made in Ashworth, including but far from limited to my office mates Max, Jon, Tom, Surabhi and Maria. I thank Mark Blaxter, Lewis Stevens and Dominik Laetsch for their instrumental help with genomics.

I thank Rob and Roz for providing such a welcoming environment in Toronto, and James, Simon and Chey for their friendship and spare room! My two summers in Toronto hold some of the fondest memories of my PhD and I have to thank everyone in the Ness and Johnson labs (as well as various others at UTM), along with many people both on and off campus downtown (Marianka, Zoë, Paul, etc.). I thank Ahmed for being such an enthusiastic collaborator. I thank Sabeeha Merchant and the Merchant lab at UC Berkeley for introducing

me to a whole new side of chlamy research (and for keeping me very busy in the last year!). I also thank David Gilbert for his tireless administrative work to enable my visit to California. Thank you to all of the wonderful people I've met at conferences who have made them such a highlight (Filipa, Henry, Emma, et al.).

I thank my family, especially my parents, sister and grandmother for all of their support, without which this would not have been possible. To all of my friends beyond academia, especially Chris, Rachel and Ryan, who always made sure I had a place to stay on my returns to Edinburgh. Finally, thank you to all of the people who've made my stay in New Zealand such a wonderful experience so far – I feel incredibly lucky to have been able to finish my thesis in the South Pacific summer, it's great to be back after all these years!

Abstract

The unicellular green alga *Chlamydomonas reinhardtii* is one of the primary model organisms in plant and algal biology. Although the species is fundamental to several research areas, including the study of photosynthesis, cilia and the cell cycle, very little is known about its evolutionary biology. Furthermore, *C. reinhardtii* research is generally limited to a single line of laboratory strains and no genomic resources exist for any closely related species. Consequently, the species has predominantly been studied in isolation, from both a population and phylogenetic perspective. In this thesis, I explore several aspects of the evolutionary genomics of *C. reinhardtii* and its closest relatives in the genus *Chlamydomonas*. I use population genomics approaches to characterise population structure across all known *C. reinhardtii* field isolates, presenting some of the first insights into the evolutionary ecology of the species. I use long read sequencing technology to produce highly contiguous genome assemblies for the three closest relatives of *C. reinhardtii*. Using these comparative resources, I describe several novel features of *Chlamydomonas* genomics, including the putative centromeric repeat. I present near complete reference assemblies for two laboratory strains of *C. reinhardtii*, characterising structural mutations that have occurred in the laboratory and revealing numerous misassemblies in previous versions. Finally, I present an exhaustively curated library of *C. reinhardtii* transposable elements and I describe a major new clade of retrotransposons present across the green lineage and animals. This collective work greatly expands our understanding of *Chlamydomonas* evolutionary genomics and is expected to be integral to the continued development of *C. reinhardtii* as a model for evolutionary biology research.

Lay Summary

Chlamydomonas reinhardtii is a microscopic single-celled alga that is found in soil. The species is well suited to work in the laboratory, where it has been used for more than 75 years to study specific features shared with either plant or animal cells. *C. reinhardtii* has also recently been used to experimentally study fundamental processes in evolutionary biology, most notably the process of mutation, which provides the basis for natural selection and adaptation. However, the species is almost exclusively studied in isolation, and we know very little about the genetics and genomics of *C. reinhardtii* in nature. We also know almost nothing about the evolution of any closely related *Chlamydomonas* species. In this thesis, I begin to address these shortfalls. I use whole genome sequencing data to characterise how wild isolates of *C. reinhardtii* are related. I use cutting edge genome sequencing approaches to produce genome sequences for three relatives of *C. reinhardtii*, which I use to explore how particular features of the *C. reinhardtii* genome evolved. I also use similar approaches to improve the genome sequence of *C. reinhardtii* itself, discovering several errors in the existing version. Finally, I present an in-depth annotation of *C. reinhardtii* transposable elements, which are selfish genetic elements that can replicate and move throughout genomes. Collectively, I use these resources to provide an evolutionary context for *C. reinhardtii*, which helps us understand its wider biology. *C. reinhardtii* and *Chlamydomonas* have potential to be developed as useful systems to study evolutionary biology, and the work and resources presented in this thesis form some of the first steps in this process.

Publications

The following manuscripts have been published and form the basis of chapters in this thesis:

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

Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. 2021. An ancient clade of *Penelope*-like retroelements with permuted domains is present in the green lineage and protists, and dominates many invertebrate genome. *Mol Biol Evol* **in press**. doi:10.1093/molbev/msab225

I contributed to the following published manuscripts during my PhD, although they do not form part of this thesis:

Craig RJ, Suh A, Wang M, Ellegren H. 2018. Natural selection beyond genes: identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*). *Mol Ecol* **27**: 476-492.

Smith DR, **Craig RJ**. 2021. Does mitochondrial DNA replication in *Chlamydomonas* require a reverse transcriptase? *New Phytol* **229**: 1192-1195.

Gallaher, D. S., **Craig, R. J.**, Ganesan, I., Purvine, S. O., McCorkle, S. R., Grimwood, J., Strenkert, D., Davidi, L., Roth, M. S., Jeffers, T., Lipton, M. S., Niyogi, K. K., Schmutz, J., Theg, S. M., Blaby-Haas, C. E. & Merchant, S. S. 2021. Widespread polycistronic gene expression in the green algal lineage. *Proc Natl Acad Sci U S A* **118**: e2017714118.

Chaux-Jukic F, O'Donnell S, **Craig RJ**, Eberhard S, Vallon O, Zhou X. 2021. Architecture and evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **49**: 7571-7587

Lopez-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* **in press.** doi:10.1093/molbev/msab140.

Table of Contents

Declaration iii

Acknowledgements iv

Abstract vi

Lay Summary vii

Publications viii

Table of Contents ix

List of Tables and Figures xiii

Abbreviations and Gene Symbols xv

Chapter 1: General Introduction 1
1.1 Thesis outline 1
1.2 *Chlamydomonas reinhardtii* and the genus *Chlamydomonas* 2
1.2.1 *Chlamydomonas reinhardtii* as a model organism 2
1.2.2 The genus *Chlamydomonas* and the wider systematics of green algae 3
1.2.3 *Chlamydomonas reinhardtii* laboratory strains 5
1.2.4 *Chlamydomonas reinhardtii* field isolates 6
1.2.5 *Chlamydomonas reinhardtii* as a study system in evolutionary biology 7
1.3 The *Chlamydomonas reinhardtii* genome 8
1.3.1 The history of the Chlamydomonas Genome Project 8
1.3.2 Genome architecture and organisation 9
1.3.3 Genetic and molecular mapping 10
1.3.4 Base composition, mutation and recombination 11
1.3.5 Genome-wide patterns of methylation 14
1.3.6 The mating type locus 14
1.3.7 Transposable elements 15
1.3.8 Structural variation between laboratory strains and field isolates 17
1.3.9 Organelle genomes 17
1.3.10 Genome assemblies of related species 18
1.4 Introduction to core methodology 18
1.4.1 Whole-genome re-sequencing and variant calling 18
1.4.2 Genome assembly using long read sequencing 21
1.4.3 Transposable element classification and annotation 23
1.5 *Chlamydomonas* in a wider evolutionary context 27

Chapter 2: Patterns of Population Structure and Complex Haplotype Sharing Among Field Isolates of the Green Alga *Chlamydomonas reinhardtii* 30
2.1 Preface 30
2.2 Abstract 31
2.3 Introduction 32

2.4 Results 33
2.4.1 Whole-genome re-sequencing of *Chlamydomonas reinhardtii* field isolates 33
2.4.2 Patterns of continental population structure 34
2.4.3 Population structure inferences from the organelle genomes 37
2.4.4 Admixture profiling and the identification of putatively introgressed genomic regions 38
2.4.5 Identity by descent sharing and patterns of local population structure 40
2.4.6 Genetic diversity within lineage, and genetic differentiation and divergence between lineages 43
2.5 Discussion 45
2.5.1 Overview 45
2.5.2 The North American biogeography of *Chlamydomonas reinhardtii* 45
2.5.3 The Japanese isolates and the wider biogeography of *Chlamydomonas reinhardtii* 47
2.5.4 Patterns of population structure and genetic diversity at a local scale 48
2.5.5 Broader perspectives on microbial biogeography and speciation 49
2.5.6 Identity by descent sharing between *Chlamydomonas reinhardtii* isolates 49
2.6 Conclusions 50
2.7 Methods 50
2.7.1 Sampling and whole-genome re-sequencing 50
2.7.2 Read mapping and variant calling 51
2.7.3 Identification of clonal isolates 52
2.7.4 Quebec isolates found not to be *Chlamydomonas reinhardtii* 52
2.7.5 Genomic site class annotations 52
2.7.6 Population structure analyses 53
2.7.7 Mitochondrial and plastid haplotype networks 54
2.7.8 Admixture profiling and the identification of putatively introgressed genomic regions 54
2.7.9 Identification of genomic tracts inherited identical by descent 54
2.7.10 Calculation of population genetics statistics within and between lineages 55
2.8 Acknowledgements 55

Chapter 3: Comparative Genomics of *Chlamydomonas* 56
3.1 Preface 56
3.2 Abstract 57
3.3 Introduction 58
3.4 Results and Discussion 59
3.4.1 The closest known relatives of *Chlamydomonas reinhardtii* 59
3.4.2 The genomes of *Chlamydomonas incerta*, *Chlamydomonas schloesseri* and *Edaphochlamys debaryana* 61
3.4.3 Phylogenomics of the core-*Reinhardtinia* and Volvocales 64
3.4.4 Conserved genome architecture and centromeric structure in *Chlamydomonas* 65
3.4.5 Gene and gene family evolution in the core-*Reinhardtinia* 70
3.4.6 Evolution of the mating type locus in *Chlamydomonas* 74
3.4.7 Alignability and estimation of neutral divergence 76
3.4.8 False positive and missing genes in the *Chlamydomonas reinhardtii* v5.6 annotation 78

3.4.9 The genomic landscape of sequence conservation in *Chlamydomonas reinhardtii* 81

3.5 Conclusions 84

3.6 Methods 85

3.6.1 Nucleic acid extraction and sequencing 85

3.6.2 *De novo* genome assembly of *Chlamydomonas incerta* 85

3.6.3 *De novo* genome assembly of *Chlamydomonas schloesseri* 87

3.6.4 *De novo* genome assembly of *Edaphochlamys debaryana* 88

3.6.5 Annotation of genes and repetitive elements 88

3.6.6 Phylogenomic analyses 89

3.6.7 General comparative genomics and synteny analyses 90

3.6.8 Gene annotation metrics and gene family evolution 90

3.6.9 Mating type locus evolution 91

3.6.10 Whole-genome alignment and estimating divergence 91

3.6.11 Identification of false positive and missing genes in *Chlamydomonas reinhardtii* 92

3.6.12 Conserved element identification and analyses 93

3.7 Acknowledgements 93

Chapter 4: The Chlamydomonas Genome Project, Version 6: Near Complete Genome Assemblies for Mating Type *Plus* and *Minus* Strains Reveals Extensive Misassembly in Version 5 and Structural Mutation in the Laboratory 94

4.1 Preface 94

4.2 Abstract 95

4.3 Introduction 96

4.4 Results and Discussion 98

4.4.1 CC-503 version 6: a long read *Chlamydomonas reinhardtii* reference genome 98

4.4.2 The version 6 assembly reveals misassemblies in version 5 99

4.4.3 Sequence context of filled gaps and the final assembly challenges 103

4.4.4 The CC-503 genome harbours major structural mutations 106

4.4.5 A reference quality assembly of CC-4532 and a complete mating type *minus* locus 109

4.4.6 The CC-4532 assembly reveals transposable element proliferation in the laboratory 111

4.4.7 Structural annotation of the version 6 assemblies 114

4.4.8 Transposable element genes and the interesting example of *CMD1* 116

4.4.9 The present and future of the Chlamydomonas Genome Project 118

4.5 Methods 120

4.5.1 Chromosomal assemblies of CC-503 and CC-4532 120

4.5.2 Identification of repetitive sequences 121

4.5.3 Comparison and lift over between assemblies 122

4.5.4 Curation of structural mutations and transposition events 122

4.5.5 Structural annotation and designation of low coding potential genes 123

4.5.6 Lift over of missing genes between assembly versions 124

4.5.7 Annotation of transposable element genes 125

Chapter 5: Transposable element annotation in *Chlamydomonas* Reveals a Major New Clade of Retrotransposons 126

5.1 Preface 126

5.2 Abstract 127

5.3 Introduction 128

5.4 Results 130

5.4.1 An exhaustively curated *Chlamydomonas reinhardtii* transposable element library 130

5.4.2 *Chlamys*: a major new clade of *Penelope*-like elements 133

5.4.3 Functional characterisation of a *Chlamys* element 140

5.4.4 *Naiad*: a metazoan clade of *Penelope*-like elements with N-terminal endonuclease 141

5.4.5 *Chlamys*-like elements, with and without endonuclease 143

5.4.6 *Hydra*: an enigmatic clade of metazoan *Penelope*-like elements with C-terminal endonuclease 145

5.4.7 An updated phylogeny of *Penelope*-like elements 145

5.5 Discussion 148

5.5.1 Diversity and recent activity of transposable elements in *Chlamydomonas reinhardtii* 148

5.5.2 *Naiad/Chlamys* is a major new clade of retrotransposons 151

5.5.3 The evolution of *Penelope*-like elements 152

5.6 Methods 155

5.6.1 Curation of transposable elements in *Chlamydomonas* 155

5.6.3 Curation of *Naiad/Chlamys* elements in other species 156

5.6.3 Phylogenetic and protein clustering analyses 156

5.6.4 Genome assembly of CC-2931 157

Chapter 6: General Discussion 158

6.1 Thesis overview 158

6.2 Evolutionary genomics resources for *Chlamydomonas* 159

6.2.1 The *Chlamydomonas reinhardtii* reference genome 159

6.2.2 The *Chlamydomonas reinhardtii* structural annotations 160

6.2.3 Population genetics resources 162

6.2.4 Comparative genomics resources 163

6.3 Trawling for *Chlamydomonas* 165

6.4 Concluding remarks 166

References 167

Appendix A: supplementary material for Chapter 1 197

Appendix B: supplementary material for Chapter 2 199

Appendix C: supplementary material for Chapter 3 206

Appendix D: supplementary material for Chapter 4 223

Appendix E: supplementary material for Chapter 5 238

List of Tables and Figures

Chapter 1

Figure 1. Producing a consensus sequence for the active DNA transposon *Tcr3* 26

Chapter 2

Figure 1. Sampling locations and years for all field isolates included in analyses 34
Figure 2. Species-wide population structure analyses 36
Figure 3. Isolation by distance in NA1 and NA2 37
Figure 4. Local ancestry profiling and putative genome-wide introgression 39
Table 1. Average genomic proportions shared identical by descent within and between samples 41
Figure 5. Identity by descent sharing in NA1 42
Figure 6. Summary of genetic diversity within *C. reinhardtii* lineages 44
Table 2. Differentiation and divergence between *C. reinhardtii* lineages 45

Chapter 3

Figure 1. Images of *Chlamydomonas* and *Edaphochlamys* species 60
Table 1. Genome assembly metrics for eight high-quality core-*Reinhardtinia* genome assemblies 62
Figure 2. Maximum likelihood phylogeny of 15 Volvocales species and three outgroups 65
Figure 3. Synteny between *C. reinhardtii* and it's close relatives 67
Figure 4. Phylogenetic relationship and centromeric clustering of *Zepp*-like elements 69
Table 2. Gene annotation metrics for core-*Reinhardtinia* species 71
Figure 5. Gene families in the core-*Reinhardtinia* 73
Figure 6. Synteny across the *C. reinhardtii* MT^{-} haplotype and inferred MT^{-} haplotypes in *C. incerta* and *C. schloesseri* 75
Figure 7. Putatively neutral divergence and genome-wide alignability across the core-*Reinhardtinia* 77
Figure 8. Coding potential analyses and false positive genes 80
Table 3. Overlap between conserved elements and *C. reinhardtii* genomic site classes 82
Figure 9. Intron lengths and overlap with conserved elements 83

Chapter 4

Table 1. Comparison of assembly metrics between reference genome versions 99
Figure 1. Overview of the CC-503 v6 assembly 100
Figure 2. Examples of v5 misassemblies and their resolution in CC-503 v6 102
Figure 3. Filled gaps and the remaining assembly challenges in CC-503 v6 105
Figure 4. Structural mutations in the CC-503 genome 107
Figure 5. The CC-4532 v6 assembly 110
Figure 6. *Gypsy-7a_cRei* copy number estimates across laboratory strains 113
Table 2. Comparison of structural annotations between reference genome versions 114
Figure 7. Browser view examples of gene models improved between v5.6 and CC-4532 v6.1 115
Figure 8. Transposable element genes and *CMD1* 117

Chapter 5

Table 1. Comparison between Repbase and updated transposable element libraries 131
Figure 1. Transposable element landscape for *C. reinhardtii* CC-1690 132
Figure 2. Reverse transcriptase phylogeny of the *Naiad/Chlamys* clade of N-terminal EN+ PLEs. 135
Figure 3. The diversity of *Chlamys, Naiad* and *Chlamys*-like elements 136
Figure 4. Alignment of *Chlamys, Naiad* and *Chlamys*-like protein sequences 137
Figure 5. Functional characterisation of *Chlamys-3_cRei* 141
Figure 6. Reverse transcriptase phylogeny of PLEs and TERTS 146
Figure 7. CLANS protein clustering of the GIY-YIG endonuclease domain 147

Abbreviations and Gene Symbols

Abbreviations (note software abbreviations are not included):

0D	zero-fold degenerate
2D	two-fold degenerate
4D	four-fold degenerate
5gmC	C^5-glyceryl-methylcytosine
5mC	C^5-methylcytosine
6mA	N^6-methyldeoxyadenosine
A-NHEJ	alternative non-homologous end joining
APE	apurinic/apyrimidinic endonuclease-like endonuclease
AS	alternative splicing
BAM	Binary Alignment/Map
BAQ	base quality score
BUSCO	Benchmarking Universal Single Copy Ortholog
C domain	centromere-proximal mating type domain
CAI	Codon Adaptation Index
CC	Chlamydomonas Centre
CCAP	Culture Collection of Algae and Protozoa
cDNA	complementary deoxyribonucleic acid
CDS	coding sequence
CE	conserved element
cenH3	centromere-specific histone H3
ChIPseq	chromatin immunoprecipitation sequencing
CIV	Chilo iridescent virus
CLiP	Chlamydomonas Library Project
C-NHEJ	classical non-homologous end joining
CNV	copy number variant
CO	crossover
CTAB	cetyltrimethylammonium bromide
CTL	C-type lectin
DIRS	*Dictyostelium* intermediate repeat sequence-like
dN	nonsynonymous divergence
dS	synonymous divergence
DNA	deoxyribonucleic acid
DSB	double strand break
EN	endonuclease
EN–	endonuclease deficient
EN+	endonuclease containing
EST	expressed sequence tag

ETS	external transcribed spacer
GAG	group-specific antigen
GFF	General Feature Format
GQ	genotype quality
GTR	general time reversible
gVCF	Genomic Variant Call File
H3K4me3	trimethylation of lysine 4 on histone H3
HAL	Hierarchical Alignment Format
HE	homing endonuclease
HR	homologous recombination
IFD	insertion in the fingers domain
INDEL	insertion/deletion
INT	integrase
IR	intron retention
I_{TE}	Index of Translation Elongation
ITS	internal transcribed spacer
JPN	Japanese lineage of C. reinhardtii
K	STRUCTURE population number
KDZ	*Kyakuja-Dileera-Zisupton*
LD	linkage disequilibrium
LINE	long interspersed nuclear element
lncRNA	long noncoding ribonucleic acid
LTR	long terminal repeat
MA	mutation accumulation
ML	maximum likelihood
MNNG	N-methyl-N’-nitro-N-nitrosoguanidine
mRNA	messenger ribonucleic acid
MT^{+}	mating type *plus*
MT^{-}	mating type *minus*
NA1	North America 1 lineage of C. reinhardtii
NA2	North America 2 lineage of C. reinhardtii
NCL	nuclear control of chloroplast gene expression-like
NCO	non-crossover
NGS	next generation sequencing
NHEJ	non-homologous end joining
NTS	nontranscribed spacer
ONT	Oxford Nanopore Technologies
ORF	open reading frame
PacBio	Pacific Biosciences
PAV	presence-absence variant
PBCV-1	*Paramecium bursaria* chlorella virus 1
PCA	principle component analysis
PCR	polymerase chain reaction

PHD	plant homeodomain finger
PKc	protein kinase catalytic domain
PLE	*Penelope*-like element
pLTR	pseudo-long terminal repeat
PROT	pepsin-like aspartate protease
R domain	rearranged mating type domain
RADseq	restriction site-associated DNA sequencing
RC	rolling circle
rDNA	ribosomal deoxyribonucleic acid
RFLP	restriction fragment length polymorphism
RH	ribonuclease H (RNAse H)
RLE	restriction-like endonuclease
rRNA	ribosomal ribonucleic acid
RNA	ribonucleic acid
RT	reverse transcriptase
SAG	Culture Collection of Algae at Göttingen University
SAM	Sequence Alignment/Map
SAP	SAF A/B, Acinus and PIAS
SECIS	selenocysteine insertion sequence
SINE	short interspersed nuclear element
SMRT	single molecule real-time
SNM	single nucleotide mutation
SNP	single nucleotide polymorphism
SRCR	scavenger receptor cysteine-rich
SV	structural variant
T domain	telomere-proximal mating type domain
TAIR	The Arabidopsis Information Resource
TE	transposable element
TERT	telomerase reverse transcriptase
TET/JBP	ten-eleven translocation/J-binding protein
TGV	*Tetrabaenaceae-Goniaceae-Volvocaceae*
TIR	terminal inverted repeat
tRNA	transfer RNA
TSD	target site duplication
TSS	transcription start site
UCE	ultraconserved element
UTEX	Culture Collection of Algae at the University of Texas at Austin
UTR	untranslated region
VCF	Variant Call File
WGA	whole-genome alignment
YR	tyrosine recombinase
ZeppL	*Zepp*-like

Gene symbols:

ALD	Plastid fructose-1,6-bisphosphate aldolase
atpB	ATP synthase subunit beta
atpE	ATP synthase epsilon chain
chlL	Light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein
CMD1	5mC-modifying enzyme
CYTC1	Mitochondrial cytochrome c1
DHC6	Dynein heavy chain 6
DHC9	Dynein heavy chain 9
DRC4	Nexin-dynein regulatory complex 4
DYX1C1	dyslexia susceptibility 1 candidate 1
EZY2	Early zygote expressed protein 2
FUS1	Fusion 1
HDH1	Histidinol dehydrogenase 1
HIL1	Heat inducible lipase 1
INT1	Integrase 1
MAT3	Retinoblastoma protein
MID	Minus dominance
MT0796	Mating type 0796
MT0828	Mating type 0828
MTA1	Mating type A1
MTD1	minus mating type differentiation 1
MTHI1	atpH mRNA stabilization factor
MTP0428	Mating type 0428
MUT6	DEAH Box RNA helicase
NIC7	Quinolinate synthetase A
NIT1	Nitrate reductase
NIT2	Nitrate assimilation regulatory protein
ODA2	Dynein heavy chain gamma
OTU2	Otubain domain putative protease 2
petA	cytochrome f precursor
PETC	Rieske iron-sulfur subunit of the cytochrome b6f complex
PF20	Protein required for flagellar central pair microtubule assembly 20
PRPL4	Chloroplast ribosomal protein L4
psbA	Photosystem II protein D1
PSBW	Photosystem II reaction center W protein
PSY	Phytoene synthase
rbcL	Ribulose bisphosphate carboxylase large chain
RPS14	Ribosomal protein S14
rrnL	23S ribosomal RNA
RSP4	Radial spoke protein 4

RSP6	Radial spoke protein 6
SPP3	SppA, protease IV, signal peptide peptidase
TGL1	Sterol esterase TGL1
THI8	Hydroxymethylpyrimidine phosphate synthase
VFL2	20 kD Calcium-Binding Protein Centrin (caltractin)
ycf4	Photosystem I assembly protein Ycf4
YPT4	GTP-binding protein ypt4
ZYS3	Zygote-specific 3

Chapter 1

General Introduction

1.1 Thesis Outline

In this thesis, I present the results of several analyses and the associated genomic resources that develop *Chlamydomonas reinhardtii* as a model for evolutionary genomics research. Using whole-genome re-sequencing data from *C. reinhardtii* field isolates, I investigate the population structure of the species. I use long read sequencing to update the *C. reinhardtii* reference genome and to produce high-quality genome assemblies for three closely related species. Using these novel and improved genome assemblies, I characterise several features of genome architecture and evolution in *C. reinhardtii* and its relatives. I perform exhaustive manual curation of transposable elements (TEs) in *C. reinhardtii* and describe a new clade of retrotransposons that are abundant in green algae and invertebrate animals, and are also present in at least a small number of plant and protist species. *C. reinhardtii* possesses several features that make it an attractive study system for evolutionary research. The work presented in this thesis substantially improves our understanding of *Chlamydomonas* genomics and evolutionary biology, and the resources presented are expected to form a basis for future research in this area. In this chapter, I introduce *C. reinhardtii* and *Chlamydomonas* as study systems and provide an overview of the *C. reinhardtii* genome. I also present detailed background information on three methodological approaches that are central to the research I have performed, namely whole-genome re-sequencing and variant calling, long read genome assembly, and TE annotation and classification. Finally, I briefly outline the wider context of the work presented in this thesis.

1.2 *Chlamydomonas reinhardtii* and the Genus *Chlamydomonas*

1.2.1 *Chlamydomonas reinhardtii* as a model organism

C. reinhardtii, commonly referred to as Chlamydomonas (without italics) or simply "chlamy" in everyday communications, is a unicellular green alga that is widely studied across several research areas. Befitting its prominence, the species has been the subject of excellent general reviews (Harris 2001; Pröschold et al. 2005; Salomé and Merchant 2019) and two editions of a dedicated textbook, the *Chlamydomonas* Sourcebook (Harris 1989; Harris 2009), and the introduction presented in this section is expanded in far greater detail in those sources. *C. reinhardtii* cells measure ~5 x 10 μm, are encased in a glycoprotein/carbohydrate cell wall and contain two cilia that primarily function in motility (including phototaxis) and cell-cell recognition during sexual reproduction. *C. reinhardtii* is haploid and facultatively sexual, dividing mitotically in resource-rich conditions and undergoing sex when resources are limited (specifically nitrogen limitation in the laboratory). During vegetative growth, multiple mitotic divisions (from one to five, most often two) occur within the cell wall of the mother cell, resulting in the release of between two and 32 daughter cells. In the sexual cycle, cells differentiate to equal-sized (i.e. isogamous) gametes of two genetically determined mating types, *plus* (MT^{+}) and *minus* (MT^{-}), and pairs of opposite mating type gametes pair and fuse to form a diploid zygote. The zygote further develops to a desiccation and frost-resistant zygospore, which undergoes meiosis to produce four vegetative cells, two of each mating type. *C. reinhardtii* has been isolated from soil in continental and temperate climates, although it is expected to also exist in temporary water bodies. Ecological knowledge of the species is severely limited (Sasso et al. 2018), however it is likely that sex and zygospore formation are a requirement of survival in adverse conditions such as drought and cold. Hasan and Ness (2020) estimated the rate of sex in *C. reinhardtii* as approximately one sexual generation per 840 asexual generations.

A now unknown species described as *Chlamydomonas pulvisculus* was first introduced by Ehrenberg (1838), establishing the genus *Chlamydomonas* (see 1.2.2), while another long since lost isolate was first described as *C. reinhardtii* by Dangeard (1888). Early cultures of other unknown *Chlamydomonas* species were used to study phototaxis and sexual reproduction, with Pascher (1918) introducing tetrad analysis as a powerful approach to study Mendelian segregation. The contemporary history of *C. reinhardtii* commenced with the proposed isolation and subsequent germination of a zygospore from a potato field near Amherst, Massachusetts in 1945 by Gilbert M. Smith, giving rise to both MT^{+} and MT^{-} haploid progeny that led to the establishment of the "laboratory strains" (1.2.3). Over the following decades, fundamental work on these strains established *C. reinhardtii* as a model organism. The species is experimentally tractable and well suited to laboratory culture; it can be grown axenically with a doubling time of 6-8 hours, the cell cycle can be synchronised, and the sexual cycle controlled. It is highly amenable to classical genetics, and its haploid state means that loss-of-function mutations are immediately scorable. Unlike land plants and other early *Chlamydomonas* models (e.g. *Chlamydomonas moewusii*), *C. reinhardtii* can use acetate as a sole carbon source, making it a powerful system in which to study photosynthesis, since knockout mutations are non-lethal. Many green algae have also maintained certain cellular features present in animal cells but lost in land plants, and most notably *C. reinhardtii* is used as a model for the biogenesis and function of cilia. The species has been used to study the cell cycle, chloroplast biology, cell metabolism, photoreceptors and a variety of other phenomena. Its role as the best studied algal species also makes *C. reinhardtii* an important reference system for algal biotechnology and industrial applications such as the production of biofuels and bioproducts (Scranton et al. 2015). Although development of a molecular toolkit proved more complicated than in many other classical models, genetic transformation is now possible, and in the last two decades the sequencing of the *C. reinhardtii* genome (1.3) has enabled many new experimental approaches, most notably high-throughput transcriptomics. Overall, its tractability as an experimental and classical genetics system has led to comparisons with other important microbial models, with *C. reinhardtii* occasionally referred to as the "green" or "photosynthetic yeast" (Goodenough 1992; Rochaix 1995).

1.2.2 The genus *Chlamydomonas* and the wider systematics of green algae

Chlamydomonas green algae belong to the phylum Chlorophyta, class Chlorophyceae, and order Volvocales (=Chlamydomonadales). The chlorophytes, together with the streptophytes (land plants + their green algal relatives), red algae and glaucophytes, are members of the Archaeplastida, a major clade of eukaryotes that evolved following the hypothesised primary endosymbiosis event that gave rise to plastids ~1.5 billion years ago (Parfrey et al. 2011). The chlorophytes and streptophytes likely diverged more than one billion years ago, and collectively form the clade Viridiplantae (Leliaert et al. 2012). The Chlorophyceae are one of the three major lineages that represent the core-Chlorophyta, together with the Trebouxiophyceae and Ulvophyceae, which likely last shared a common ancestor in the Neoproterozoic era >700 million years ago (Del Cortona et al. 2020).

The Volvocales is a highly diverse order of mostly freshwater and terrestrial species, including both unicellular and multicellular, motile and non-motile, and heterothallic and homothallic species, as well as a number of extremophiles (e.g. the halotolerant *Dunaliella salina*, acidophile *Chlamydomonas eustigma* and psychrophile *Chlamydomonas nivalis*). Alongside *Chlamydomonas*, one of the most recognisable groups in the Volvocales are the colonial volvocine algae, which includes species spanning a large range of organismal complexities, from the four-celled isogamous *Tetrabaena socialis* to the multicellular anisogamous *Volvox carteri*. The group have therefore attracted significant attention as models to study major evolutionary transitions, including the evolution of multicellularity, anisogamy, and the differentiation of soma and germ line (Umen 2020). The volvocine algae are also known as the TGV clade, which refers to the three major constituent families, the Tetrabaenaceae, Goniaceae and Volvocaceae.

The traditional genus *Chlamydomonas* contains more than 600 described species of biflagellate unicellular algae, although it was already considered to be an artificial assemblage by Ettl (1976). The systematics of *Chlamydomonas* and its relatives have since been characterised using marker gene phylogenies, most commonly the 18S ribosomal RNA (rRNA) gene. Pröschold et al. (2001) confirmed that *Chlamydomonas* is highly polyphyletic and began the arduous task of revising the genus and transferring species to new genera. Nakada et al. (2008) defined several major clades within the Volvocales (e.g. *Reinhardtinia*, *Moewusinia*, *Dunaliellinia*), several of which included species described as *Chlamydomonas*. *C. reinhardtii* belongs to a subclade of the *Reinhardtinia* termed the core-*Reinhardtinia*, which also includes the TGV clade and many species described as *Chlamydomonas* (e.g. *Chlamydomonas sphaeroides*, *Chlamydomonas zebra*, *Chlamydomonas cribum*) and *Vitreochlamys* (Nakada et al. 2016). The wider *Reinhardtinia* includes several other *Chlamydomonas* species and representatives of other genera such as *Neochlorosarcina* and the nonphotosynthetic *Polytomella*. Thus, *C. reinhardtii* is more closely related to the TGV clade than the vast majority of described *Chlamydomonas* species, and *Chlamydomonas* as a traditional genus spans several hundred million years of evolution.

C. reinhardtii is the type species of *Chlamydomonas*, and therefore *C. reinhardtii* and its immediate relatives comprise the monophyletic genus *Chlamydomonas* (i.e. *sensu stricto*) (Pröschold and Silva 2007). However, while the more conspicuous TGV clade contains ~50 described species, the unicellular lineage that includes *C. reinhardtii* contains only two other described species, *Chlamydomonas incerta* and *Chlamydomonas schloesseri* (Pröschold et al. 2005; Pröschold et al. 2018). Work continues to reassign more distantly related *Chlamydomonas* species in the core-*Reinhardtinia* to new genera, for example, *Chlamydomonas debaryana*, one of the most well-sampled species, was recently renamed *Edaphochlamys debaryana* (Pröschold et al. 2018). Marker gene phylogenies have been less successful at characterising fine-scale phylogenetic relationships, and it is unclear whether unicellular species such as *E. debaryana* are more closely related to *Chlamydomonas* (*sensu stricto*), the TGV clade, or are outgroups to both. Although *Chlamydomonas* (*sensu stricto*) currently contains just three species, it is likely that many undescribed close relatives exist. A Korean freshwater isolate was described by Hong et al. (2013) that exhibited a single nucleotide substitution relative to *C. reinhardtii* in the 18S rRNA gene (for comparison the closest known relative *C. incerta* contains two substitutions), although unfortunately this isolate was lost to culture. Khaw et al. (2020) isolated a new strain from Malaysia that was identified as *C. reinhardtii* based on 100% sequence similarity between 18S rRNA genes, and a preliminary analysis based on a *de novo* transcriptome assembly has suggested that this isolate represents a new species (Craig, Balogun & Ness, unpublished). A small number of 18S rRNA sequences from undescribed species more similar to the 18S rRNA sequence of *C. reinhardtii* than to those of the TGV clade have been uploaded to GenBank, namely isolate UMT-B14 (accession MN879273.1), strain rsemsu Chlam-15/11 (KU9263335.1) and strain hoo2 (MH699043.1). Fossil calibrated molecular clock analyses have estimated that the TGV clade last shared a common ancestor with unicellular species ~230 million years ago (Herron et al. 2009), and *Chlamydomonas* (*sensu stricto*) is therefore expected to have evolved in approximately the last one hundred million years (3.4.7).

1.2.3 *Chlamydomonas reinhardtii* laboratory strains

With few exceptions, research in *C. reinhardtii* is performed on a large collection of clonally maintained cultures that are collectively known as the standard laboratory strains. As introduced in 1.2.1, these strains are all thought to be descended from a single diploid zygospore isolated in Massachusetts, 1945. However, the early history of the strains is complicated and often poorly documented. The traditional model splits the laboratory strains into three sublines based on the distribution of pairs of strains of opposite mating type to different research groups in the 1950s, namely the Sager, Cambridge and Ebersold/Levine sublines (Pröschold et al. 2005; Harris 2009). These strains have been maintained as clones in various laboratories and culture centres for approaching 70 years and are sometimes referred to as "wild type" laboratory strains, although some lines have acquired mutations, with the loss of the ability of the Ebersold/Levine strains to utilise nitrate being the most well-known example. Many additional strains have been produced by crosses between the wild type laboratory strains and their progeny. Further confusing matters, strains can be referred to by historical names (e.g. 137c, 21 gr), some of which relate to mutant phenotypes (e.g. *y1*, *cw92*), or by strain IDs from culture collections. In contemporary literature, strains are most often referred to by their "CC number" as catalogued by the Chlamydomonas Resource Centre (www.chlamycollection.org) e.g. the Sager subline strains 21 gr and *y1* are referred to as CC-1690 and CC-1691, respectively. However, as strains are frequently maintained in multiple collections, multiple synonymous IDs can be encountered e.g. the Cambridge subline strain CC-1010 is also maintained as UTEX 90 (Culture Collection of Algae at the University of Texas at Austin), SAG 11-32b (Culture Collection of Algae at Göttingen University) and CCAP 11/32A (Culture Collection of Algae and Protozoa, Scottish Association for Marine Science). Table S1 provides information on all laboratory strains referred to in this thesis.

To better resolve the genetic relationships between laboratory strains, Gallaher et al. (2015) performed whole-genome re-sequencing on 39 strains, including both the MT^{+} and MT^{-} representatives of the wild type sublines. In line with the expectation that the strains are derived from a single zygospore, they reported that the genomes of all strains are comprised of two alternative haplotypes divergent at ~2% of sites. This pairwise divergence is approximately equal to genetic diversity amongst isolates sampled from the same site (1.2.4, 2.4.6), and the two haplotypes presumably represent the genetic differences between the parental haploid individuals that once existed at the sampling site. However, one haplotype was found to be dominant, with all strains sharing at least ~75% of their genomes (and usually far higher) identical by descent in any pairwise comparison. Furthermore, the wild type strains did not correspond to four hypothetical meiotic products. This result implies that at least some of the wild type strains are the product of additional laboratory crosses that were performed in the small number of years between the original isolation and the foundation of the sublines. Moreover, as a result it was not possible to reconstruct the ancestral parental haplotypes from the wild type strains. Haplotypes were therefore arbitrarily defined relative to the reference genome (strain CC-503, 1.3.1), with the entirety of the reference genome defined as haplotype 1, and any region featuring the alternative haplotype relative to the reference genome in any other strain defined as haplotype 2. Gallaher et al. (2015) also reported several phenotypic differences between strains, and generally concluded that researchers should not consider laboratory strains to be either isogenic or experimentally equivalent.

1.2.4 *Chlamydomonas reinhardtii* field isolates

The history of *C. reinhardtii* field sampling is unfortunately brief, which mostly results from a lack of ecological knowledge and the difficulty in distinguishing *C. reinhardtii* from other *Chlamydomonas* species. This is exemplified by the field isolate CC-1373, which was sampled by G. M. Smith from a tobacco field in Massachusetts at a similar time as the laboratory strains (1.2.3). Although it is interfertile with laboratory strains and has since been genetically confirmed as *C. reinhardtii* (Coleman and Mai 1997), CC-1373 was long considered as the separate species *Chlamydomonas smithii* based on morphological differences (Hoshaw and Ettl 1966). Subsequent sampling efforts have generally relied on successful mating with laboratory strains to identify new isolates of *C. reinhardtii*. Two isolates, known as S1 D2 (CC-2290) and S2 C5 (CC-1952), were isolated from a soil sample in Minnesota by Gross et al. (1988). Although the two strains are thought to be clones (Jang and Ehrenreich 2012), they are by far the most well-studied field isolates, and crosses between CC-2290/CC-1952 and laboratory strains were the basis of the *C. reinhardtii* molecular maps (Kathir et al. 2003; Rymarquis et al. 2005). Two isolates from Pennsylvania and one from Florida were sampled by Spanier et al. (1992), while Elizabeth H. Harris isolated three strains from garden soil in North Carolina in 1991. Graham Bell's research group isolated more than 20 new strains from two fields ~80 km apart in Quebec in 1993 and 1994, four of which were described by Sack et al. (1994). Most recently, Nakada et al. (2010) isolated two new strains from a rice paddy in Kagoshima, Japan, representing the first isolates of *C. reinhardtii* from outside N. America. Information on all *C. reinhardtii* field isolates is provided in Appendix B, Table S1. The sampling history of close relatives of *C. reinhardtii* is even more limited and is introduced in 3.4.1.

Two studies by Jang and Ehrenreich (2012) and Flowers et al. (2015) have performed whole-genome re-sequencing on subsets of the N. American field isolates. These studies revealed two major results. First, Flowers et al. (2015) estimated species-wide genetic diversity (π) to be ~2.8% genome-wide, a value that is amongst the highest reported in eukaryotes (Leffler et al. 2012). Substantial variation in protein-coding sequence was also reported, with nonsynonymous diversity estimated as 0.69% and over a thousand genes predicted to contain a loss-of-function mutation in at least one strain. Second, both studies reported the presence of genetic population structure that corresponded to the sampling locations of isolates. This principally divided isolates from the North East (laboratory strains and CC-1373 from Massachusetts, and the Quebec isolates) from those sampled from further south and west in the USA. As population structure can confound population genetics inferences (Städler et al. 2009), the ~20 Quebec isolates likely represent the best opportunity to estimate and study fundamental population genetics metrics in the species. Unfortunately, only four of these isolates were sequenced by the two studies. I further explore the population structure of *C. reinhardtii* using data from all known field isolates in Chapter 2.

1.2.5 *Chlamydomonas reinhardtii* as a study system in evolutionary biology

Despite its extensive use in other fields, *C. reinhardtii* has not been widely studied from an evolutionary perspective. A notable exception is the field of experimental evolution, to which *C. reinhardtii* is well suited (fast generation times, control of the sexual cycle, etc.). To give a non-exhaustive list of examples, experiments using *C. reinhardtii* have been performed to investigate the role of sex in adaptation to novel environments (Colegrave 2002; Colegrave et al. 2002; Kaltz and Bell 2002), evolution in fluctuating environments (Reboud and Bell 1997; Kassen and Bell 1998), adaptation to elevated levels of CO_2 (Collins and Bell 2004; Collins et al. 2006) and salinity (Lachapelle et al. 2015), the evolution of herbicide resistance (Vogwill et al. 2012; Lagator et al. 2013), and the evolution of multicellularity (Ratcliff et al. 2013; Herron et al. 2019). Many of the additional applications of *C. reinhardtii* in evolutionary biology have also stemmed from its experimental amenability. The species is an excellent system for studying *de novo* mutation in mutation accumulation (MA) experiments (Ness et al. 2012; Sung et al. 2012; Ness et al. 2015; Ness et al. 2016), since its ~111 Mb genome (1.3) provides a far larger target for mutation relative to other experimental models such as yeasts. The relationship between fitness and the number of *de novo* mutations has been explored (Kraemer et al. 2017), and most recently further experimental work on existing MA lines enabled Böndel et al. (2019) to infer the distribution of fitness effects of *de novo* mutations in the species. Whole-genome re-sequencing of *C. reinhardtii* tetrads was performed by Liu et al. (2018) to experimentally assess GC-biased gene conversion (1.3.4).

Beyond experimental work, analyses at the population and between-species levels have been severely impeded by a lack of biological samples and genomic resources. As reported in 1.2.4, Jang and Ehrenreich (2012) and Flowers et al. (2015) characterised general features of genetic diversity in the species, although their analyses were performed on subsets of the available field isolates that exhibited substantial population structure. More recently, Hasan and Ness (2020) assessed recombination rate variation in *C. reinhardtii* (1.3.4) using the Quebec field isolates as a population dataset. The mating type locus has been studied in more detail, and De Hoff et al. (2013) and Hasan et al. (2019) used population data to characterise genetic differentiation between the *MT^+ and MT^-* haplotypes (1.3.6). Smith and Lee (2008) performed a comparative analysis of genetic diversity between the nuclear and mitochondrial genomes.

Broad-scale comparative genomics analyses have been performed between *C. reinhardtii* and members of the TGV clade, most notably *V. carteri* (Ferris et al. 2010; Prochnik et al. 2010). However, as discussed in Chapter 3, the divergence between *C. reinhardtii* and the volvocine algae is too great to perform nucleotide-level alignment, limiting the scope of analyses. Ferris et al. (1997) sequenced the sex determining gene *MID* (*MINUS DOMINANCE*) in *C. incerta*, reporting that it exhibited substantial nonsynonymous (~13%) and synonymous (~58%) divergence relative to the *C. reinhardtii* ortholog. Liss et al. (1997) reported an average divergence of ~52% between *C. reinhardtii* and *C. incerta* based on the sequencing of three introns. Popescu et al. (2006) used a *C. incerta* expressed sequence tag (EST) library to quantify divergence relative to *C. reinhardtii* and explore patterns of synonymous codon usage (1.3.4) between the species, estimating synonymous divergence as ~37% from a set of 67 orthologs. Finally, Popescu and Lee (2007) and Hua et al. (2012) performed comparative analyses of the mitochondrial genomes and of several plastid genes between *C. reinhardtii* and *C. incerta*, reporting that synonymous divergence in the organelle genomes was comparable to that found in nuclear genes. With the exception of phylogenetic marker genes, no comparative genetics or genomics analyses have been performed using *C. schloesseri*.

1.3 The *Chlamydomonas reinhardtii* Genome

1.3.1 The history of the Chlamydomonas Genome Project

Efforts to sequence, assemble and annotate a *C. reinhardtii* reference genome began in the early 2000s. A cell wall-less mutant, CC-503 (=*cw92*) was chosen as the reference strain, since this phenotype simplified DNA extraction and enabled the high yield requirements of library preparation at the time to be met. CC-503 was derived from the wild type *MT*$^+$ laboratory strain CC-125 (=137c) of the Ebersold/Levine subline (1.2.3), with the cell wall-less phenotype induced by mutagenesis (Hyams and Davies 1972). Preceded by two preliminary versions (v1 and v2, see Grossman et al. (2003)), the first high-quality draft assembly (v3) was assembled from 13x coverage of Sanger sequenced reads from plasmids, fosmids and one BAC library (Merchant et al. 2007). The v3 assembly consisted of more than 1,500 scaffolds, was 120.2 Mb and contained an estimated 12.5% gaps. This was quickly followed by the release of the 112.3 Mb v4 assembly in 2008, a complete reassembly that was assembled as 17 chromosomes and 71 unplaced scaffolds via comparison to molecular mapping data (1.3.3). The v4 assembly also incorporated targeted Sanger sequencing of assembly gaps, reducing the estimated proportion of gaps to 7.5%. With the onset of next-generation sequencing (NGS), the v5 assembly released in 2012 utilised both 454 and additional Sanger sequencing to target all remaining gaps, successfully filling approximately half of those present in v4. At 111.1 Mb, with ~3.7% gaps, 37 unplaced scaffolds and a contig-level N50 of ~220 kb (i.e. half of the assembled genome is present on contigs of 220 kb or greater), the v5 assembly has been the most long-standing and stable release to date (Blaby et al. 2014).

As with the genome assembly, the structural annotations, which define the genomic coordinates of genes and the proteins they encode, have also undergone several rounds of improvement (Blaby et al. 2014; Blaby and Blaby-Haas 2017). Annotations performed on the v3 assembly combined *ab initio* gene prediction with Sanger sequenced ESTs and complete complementary DNAs (cDNAs), resulting in the annotation of 15,143 protein-coding genes. The final annotation version (v4.3) based on the v4 assembly used the AUGUSTUS gene prediction algorithm (Stanke et al. 2008) in combination with 454 sequenced ESTs and protein homology to the newly available *V. carteri* gene annotations (Prochnik et al. 2010), resulting in 17,114 genes. The annotations performed for v5 made full use of NGS technology via the incorporation of over one billion RNA-seq reads, resulting in several major changes to gene models (e.g. splitting or fusion of gene models (Blaby and Blaby-Haas 2017)). The current v5 annotation (v5.6) contains 17,741 protein-coding gene models and 1,785 alternative transcripts.

The *C. reinhardtii* genome assembly and annotation are maintained at Phytozome (Goodstein et al. 2012), which also provides a genome browser with several functional data tracts (e.g. gene expression). Although these resources are fundamental to almost all modern aspects of *C. reinhardtii* research, they have not been updated using the latest advances in sequencing technology (1.4.2). The scale of improvement that is now possible was recently demonstrated by a *de novo* assembly of the laboratory strain CC-1690 (=21 gr) based on an ultra-long read Nanopore dataset (Liu et al. 2019), in which the 17 chromosomes were assembled as just 21 contigs (O'Donnell et al. 2020). In Chapter 4, I present Version 6 of the Chlamydomonas Genome Project.

1.3.2 Genome architecture and organisation

The *C. reinhardtii* nuclear genome is approximately 111 Mb in length, GC-rich (64.1% genome-wide) and arranged on 17 chromosomes ranging from 3.8 Mb to 9.8 Mb in length. Despite its relatively large size, the genome is highly compact, with genic sequence comprising ~85% of the v5 assembly and a median intergenic distance of only 134 bp between genes. The high gene density is largely a result of an unusual intron-richness. Intronic sequence comprises approximately one third of the genome and each gene contains eight introns on average, a number that is more similar to the human genome than species with comparable genome sizes (Merchant et al. 2007). Introns are also unusually long, with a median length of 229 bp. Indeed, the short introns of 60-110 bp that typically dominate the intron length distributions of smaller eukaryotic genomes such as *Drosophila melanogaster* and *Arabidopsis thaliana* constitute only ~5% of introns in *C. reinhardtii* (Merchant et al. 2007). The genome has been reported to contain a relatively uniform density of genes, TEs and simple repeats (Merchant et al. 2007).

Most genomics research in *C. reinhardtii* has focussed on gene function, and there have been very few studies characterising specific features of genome architecture or exploring their evolution. In the genome paper, Merchant et al. (2007) provided a great level of detail, although unfortunately the v3 assembly was not entirely assembled to chromosomes and was ~9 Mb larger than the current v5 assembly. Furthermore, many of the fine-scale annotations performed have not been carried over to subsequent versions and are not readily available. For example, Merchant et al. (2007) annotated 259 transfer RNA (tRNA) genes which were reported to be highly clustered, although the locations of these genes are not annotated in v5 and no assessment of their chromosomal distribution has been reported. Even centromeres, one of the most fundamental genomic features that define chromosome arms, have proved difficult to map precisely. Approximate centromere locations were known from molecular mapping (Preuss and Mets 2002), but only a recent study by Lin et al. (2018) defined putative centromeric coordinates. They reported that regions known to be tightly linked to centromeres on 15 of the 17 chromosomes were characterised by the presence of 200-800 kb stretches of sequence containing multiple genes encoding proteins featuring reverse transcriptase (RT) domains. They also performed additional crosses between a laboratory strain and the Minnesota field isolate CC-1952 (1.2.4), and reported very little recombination between the putative centromeres and markers tightly linked to these regions. These results suggest that centromeres in *C. reinhardtii* likely contain retrotransposons as a major constituent, as is the case in many taxa including the distantly related green algae *Coccomyxa subellipsoidea* (Blanc et al. 2012) and *Chromochloris zofingiensis* (Roth et al. 2017), although these sequences have not yet been characterised. In Chapters 3 and 4, I aim to extend our knowledge of genome architecture in *C. reinhardtii*, with a specific focus on centromeres, intron evolution and repeat content.

1.3.3 Genetic and molecular mapping

Cytological approaches have proved challenging in *C. reinhardtii* due to the difficulty of visualising meiotic chromosomes in the zygote, and there was much historical debate over the chromosome number of the species (Harris 2009). By applying electron microscopy to germinating zygotes, Storms and Hastings (1977) estimated chromosome number to be 18-20. More recently, Aoyama et al. (2008) fluorescently stained DNA in zygotes and observed 18 chromosomes, with the largest being 3-6 times longer than the shortest.

In contrast, as introduced in 1.2.1, *C. reinhardtii* is highly amenable to classical genetics research and has been the focus of crossing experiments since its isolation. Early genetic maps were built by performing tetrad analysis on crosses between opposite mating type laboratory strains (1.2.3) that carried scorable mutations. Examples of mutant phenotypes included auxotrophy, immobility and drug resistance (Eversole 1956; Hastings et al. 1965). By observing the independent segregation or co-segregation of many mutant phenotypes relative to one another, linkage groups were established. The process of determining linkage of new mutations to existing mutations and their associated linkage groups was greatly facilitated by the creation of test strains carrying multiple mutant phenotypes (Smyth et al. 1975). This work led to the establishment of 17 accepted linkage groups termed I to XIX, since two pairs of historically named groups were combined (XII with XIII, and XVI with XVII) (Dutcher et al. 1991).

Early attempts at molecular mapping utilised restriction fragment length polymorphisms (RFLPs) between laboratory strains and the Massachusetts field isolate CC-1373 (i.e. *C. smithii*, 1.2.4) to map cloned flagellar genes (Ranum et al. 1988). Far more extensive mapping was performed based on crosses between the laboratory strain CC-1690 and the Minnesota field isolate CC-1952, culminating in the molecular map of Kathir et al. (2003). This map featured 264 markers scored by either RFLPs or PCR, utilising the extensive polymorphism present between the two strains. The molecular map was successfully anchored to the 17 linkage groups of the genetic map using markers that corresponded to genes underlying genetically mapped mutant phenotypes, which were available for most linkage groups. The total map length was 1,025 cM, and any position in the genome was estimated to be ~2 cM from a molecular marker, on average. Rymarquis et al. (2005) developed additional markers and produced an improved molecular map of 506 markers, which was subsequently used to scaffold the v4 genome assembly to chromosomes (Blaby et al. (2014), 1.3.1). Evidence from genetic mapping, molecular mapping and genome assembly supports the existence of 17 chromosomes in *C. reinhardtii*, and the discrepancy with cytological estimates of 18 or more chromosomes remains unclear.

1.3.4 Base composition, mutation and recombination

The overall GC content of the *C. reinhardtii* genome is 64.1%. Furthermore, the high GC content is relatively uniform; if the genome is split into nonoverlapping 20 kb windows, GC content in 98% of genome falls between 58.5% and 69.6%. However, when considering site classes independently, variation in GC content is more pronounced. GC content is highest in coding sequence (CDS) (70.3%), lowest in 5' untranslated regions (UTRs) (54.7%) and 3' UTRs (58.3%), and intermediate in intronic (62.0%) and intergenic (61.3%) sequences. Breaking down CDS by site degeneracy, GC content is 79.5% in four-fold degenerate (4D) sites, 85.6% in two-fold degenerate (2D) sites and 64.1% in zero-fold degenerate (0D) sites. GC content is expected to evolve under the influences of mutation, GC-biased gene conversion, selection and drift, and some of the observed variation can be attributed to these forces in *C. reinhardtii*.

C. reinhardtii has been the focus of several MA experiments (1.2.5), in which clonal lines are maintained for many generations at a minimal effective population size, allowing mutations to accumulate across the genome (Halligan and Keightley 2009; Katju and Bergthorsson 2019). Ness et al. (2015) performed an experiment with 85 MA lines that identified 6,843 mutations by whole-genome re-sequencing, which was the largest single set of mutations described from any study at the time. They estimated an overall mutation rate (μ) of 11.5 x 10^{-10} per site per generation, and a single nucleotide mutation rate (μ_{SNM}) of 9.63 x 10^{-10}. 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 mutation spectrum. AT-biased mutation spectra have been observed across a wide range of eukaryotes and prokaryotes (Hershberg and Petrov 2010; Zhu et al. 2014; Krasovec et al. 2017), and may be near-universal. The estimated equilibrium GC content under the inferred mutation spectrum is 29%, implying that GC-biased gene conversion or selection, or both, play a major role in shaping GC content in *C. reinhardtii*. By modelling the genomic properties of their mutation dataset, Ness et al. (2015) calculated the probability of mutation (or mutability) at all sites in the genome. They estimated that mutability was highest in 5' and 3' UTRs (μ = 1.37 x 10^{-9}) and lowest in 0D and 4D sites (μ = 7.92 x 10^{-10}).

Recombination can directly influence GC content via GC-biased gene conversion, a process in which GC/AT heterozygous sites near the double strand breaks (DSBs) that initiate recombination are preferentially converted to GC over AT (Duret and Galtier 2009). GC-biased gene conversion is associated with both crossover (CO) and non-crossover (NCO) recombination events and is distinguished by the non-reciprocal exchange of parental alleles. Additionally, if GC content is under selection, then we may expect a positive correlation between recombination rate and GC content due to increased selection efficacy (Muller 1964; Hill and Robertson 1968; Felsenstein 1974). Liu et al. (2018) performed whole-genome re-sequencing on 21 tetrads from two crosses (between two Quebec field isolates and between a laboratory strain and a Quebec field isolate) and observed 24.4 COs per tetrad per meiosis, or ~1.4 COs per chromosome (equivalent to ~12 cM/Mb). Hasan and Ness (2020) estimated the population recombination rate (ρ) and linkage disequilibrium (LD) based on whole-genome re-sequencing of the Quebec field isolates (1.2.4). They reported a positive correlation between ρ and the CO rate as estimated by Liu et al. (2018), and a positive correlation between CO rate and genetic diversity (π), consistent with stronger effects of selection at linked sites in lower recombination regions. They found that mean ρ per chromosome was negatively correlated with chromosome length, in line with the expectation that shorter chromosomes experience higher recombination rates per bp due to the requirement of at least one CO per meiosis. LD was reported to decay to baseline levels between single nucleotide polymorphisms (SNPs) at distances of approximately 10 kb or less. Recombination rates were found to be highest in intergenic regions (specifically in regions immediately upstream of genes in longer intergenic tracts) and CDS, and lowest in UTRs. It is possible that a lower rate of recombination in UTRs may contribute to their low GC content. Flowers et al. (2015) reported that LD was elevated towards the ends of chromosome arms, which coincided with reduced genetic diversity in these regions.

Considering gene conversion, Liu et al. (2018) found that on average 0.0043% of SNPs were converted per tetrad per meiosis in *C. reinhardtii*, and that the rate of gene conversion at COs was 13x higher than at NCOs. This value compared to 1.9% of SNPs in similar crosses performed in *S. cerevisiae*. The lower value in *C. reinhardtii* relative to yeast was attributed to a lower recombination rate and shorter gene conversion tracts, which were estimated to have a median length of 364 bp for COs and 73 bp for NCOs (1,841 bp and 1,681 bp, respectively, in yeast). Of most interest, they reported no significant GC bias at COs, but a strong GC bias of 68.6% at NCOs (no significant bias was observed for either event class in yeast). Both Liu et al. (2018) and Hasan and Ness (2020) reported weak but significant positive corelations between their measures of recombination rate and GC content at local scales (2-50 kb), while only the correlation between NCOs and GC content was significant at broader scales (100-200 kb). Given that the GC-biased NCO gene conversion events appear to be far more uncommon than the unbiased CO events, it remains to be seen how strong an evolutionary force GC-biased gene conversion is in *C. reinhardtii*.

Considering selection, the elevated GC content of 2D and 4D sites can at least partially be attributed to selection acting on synonymous codon usage (i.e. translational selection, see

Bulmer (1991), Rocha (2004), Plotkin and Kudla (2011), Hanson and Coller (2018), and references therein). Naya et al. (2001) defined a set of optimal codons in *C. reinhardtii* 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 contained 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 classical property of translational selection, finding a positive relationship between optimal codons and the gene copy number and abundance of the tRNAs needed to decode them. Barahimipour et al. (2015) showed experimentally that optimising codon usage in a transgene resulted in higher translational efficiency and mRNA stability. Popescu et al. (2006) found that genes evolving under increased translational selection exhibited reduced synonymous divergence between *C. reinhardtii* and *C. incerta*. One implication of this is that synonymous sites are unlikely to provide an unbiased estimate of neutral diversity and divergence in *C. reinhardtii.* Unfortunately, there has been no systematic analysis of codon usage in *C. reinhardtii* since the genome was published, and the relationship between recombination rate and translational selection has not been explored.

Finally, it is possible that the *C. reinhardtii* genome is evolving under selection for higher GC content in general. Selection is likely required to explain GC content variation in bacteria in the face of AT-biased mutation spectra (Hershberg and Petrov 2010; Hildebrand et al. 2010), although the selective advantage of higher GC content remains unclear. Weissman et al. (2019) developed a hypothesis that linked selection for higher GC content to DNA repair in prokaryotes. They noted that bacterial species with high GC contents were associated with certain environments that induce higher rates of DNA damage and DSBs (e.g. soil microbes due to desiccation and spore formation). They reported a positive associated between GC content and the presence of the non-homologous end joining (NHEJ) DSB repair machinery, which may be favoured in species that experience many DSBs relative to the slower and more accurate homologous recombination (HR) pathway. Based on these observations, they hypothesised that higher GC content may increase the efficiency of NHEJ (and other similar repair pathways) and therefore provide a selective advantage for GC over AT alleles. Although this related specifically to prokaryotes, it is a particularly attractive hypothesis given that *C. reinhardtii* is also found in soil and that HR occurs at very low rates in the species (Zorin et al. 2005). Interestingly, GC content is thought to negatively correlate with organismal complexity in the volvocine algae (Hanschen et al. 2016), and it would be interesting to address whether multicellular species such as *V. carteri* have lower effective population sizes (or alternatively experience less DNA damage or have higher rates of HR). Indeed, genetic diversity is approximately six times lower in *V. carteri* than *C. reinhardtii* (Smith and Lee 2010), suggesting that this may be the case. Regardless of the mechanism, if selection is acting to increase GC content in *C. reinhardtii*, this has several evolutionary implications, including that no site in the genome may be truly evolving under neutrality. Addressing the relative roles of selection and GC-biased gene conversion to GC content evolution is expected to be challenging, given that a positive association with recombination would be expected for both forces. It also remains to be seen why GC content does not correlate with recombination at broad scales.

1.3.5 Genome-wide patterns of methylation

Both cytosine and adenine methylation have been characterised in *C. reinhardtii*. 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%. Salomé and Merchant (2019) hypothesised that many of these regions may coincide with the centromeres, although this has not yet been tested. Using the highly-contiguous CC-1690 assembly (1.3.1), Chaux-Jukic et al. (2021) reported hypermethylation of subtelomeric regions based on identifying 5mC directly from Nanopore reads, which are mappable in far more repetitive regions than short read bisulfite sequencing data. Although experimental support is variable (Lopez et al. 2015), these results potentially suggest a role for CG methylation in transcriptional silencing, as is thought to be the case in *V. carteri* (Babinger et al. 2007). Recently, the novel base modification C^5-glyceryl-methylcytosine (5gmC) was discovered in *C. reinhardtii* (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 catalysed by TET/JBP (ten-eleven translocation/J-binding protein) enzymes (Aravind et al. 2019).

Adenine methylation (N^6-methyldeoxyadenosine, 6mA) in *C. reinhardtii* has been characterised in remarkably fine detail by Fu et al. (2015). 6mA, which is centred on AT dinucleotides, is highly localised at promoters and forms a bimodal distribution with peaks either side of the transcription start site (TSS). The 6mA enrichment within the peaks shows a periodicity of 130-140 bp, which corresponds precisely to the 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.

1.3.6 The mating type locus

As introduced in 1.2.1, *C. reinhardtii* is heterothallic and the mating type of vegetative haploid cells is genetically determined. The 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). Ferris et al. (2010) sequenced the MT^- locus of the Minnesota field isolate CC-2290 to facilitate a direct comparison to the MT^+ locus present in the reference genome. 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 a small number of mating type-specific autosomal insertions and an ~160 kb tandemly repeated region known as the "16 kb repeats" (De Hoff et al. (2013) and 4.4.5). Mating type is determined in a dominant manner by the presence of the MT^--specific gene *MID* (Ferris and Goodenough 1997). However, the inactivation of *MID* alone is not sufficient to produce viable MT^+ gametes, since functional copies of MT^+-specific genes such as *FUS1* (*FUSION 1*) are required (Ferris et al. 1996). Figure S1 shows the domain and gene organisation of the MT^+ and MT^- loci.

From an evolutionary perspective, the mating type locus is the best studied region of the *C. reinhardtii* genome. CO recombination is suppressed across the R domain, although the shared genes of MT^+ and MT^- (i.e. gametologs) have not undergone significant differentiation as a result of gene conversion (De Hoff et al. 2013; Hasan et al. 2019). Despite elevated LD across the R domain, mating type genes with gametologs do not appear to be evolving under reduced selection efficacy relative to autosomal genes, suggesting that gene conversion is sufficient to avert the degenerative effects of CO suppression (Hasan et al. 2019). Conversely, mating type-specific genes have long been known to possess unusual characteristics, and both *MID* and *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), in line with the reduced selection efficacy and lack of GC-biased gene conversion resulting from the absence of both CO and NCO recombination (1.3.4).

Briefly, the mating type loci of volvocine algae, which are homologous to that of *C. reinhardtii*, have also been studied from a comparative genomics perspective. In *V. carteri*, the mating type locus is five times larger than that of *C. reinhardtii*, contains more mating type-specific genes, a higher repeat content, and exhibits substantially higher genetic differentiation between gametologs (Ferris et al. 2010). These findings are in line with theoretical expectations of increased mating type complexity resulting from the evolution of anisogamy and UV sex chromosomes (Charlesworth 1978; Bachtrog et al. 2011; Immler and Otto 2015). However, recent mating type sequences from the isogamous species *Gonium pectorale* and *Yamagishiella unicocca*, and the anisogamous *Eudorina* sp., have revealed a more complex picture (Hamaji et al. 2016b; Hamaji et al. 2018). While the mating type loci of *G. pectorale* and *Y. unicocca* are comparable in size to that of *C. reinhardtii*, the mating type locus of *Eudorina* sp. is comparatively tiny (R domain 90 kb in female and 7 kb in male, N.B. female is analogous with MT^+ and male with MT^-), suggesting that anisogamy can evolve without increased complexity of the mating type locus (Hamaji et al. 2018). These studies also observed very little synteny across mating type loci of the different species, implying recurrent haplotype reformation and a lack of evolutionary 'strata'. No mating type loci of species more closely related to *C. reinhardtii* have been sequenced, a situation that is partially addressed in Chapter 3.

1.3.7 Transposable elements

C. reinhardtii has a rich but understated role in the history of TE research. Given its phylogenetic distance from other model species, TEs discovered in *C. reinhardtii* have often been amongst the first representatives of entirely new TE clades (see 1.4.3 for information on

TE diversity and classification). In the pre-genome era, a small number of active TEs were experimentally characterised and fully or partially sequenced. Day et al. (1988) and Day and Rochaix (1991) described *TOC1*, an unusual 5.7 kb retrotransposon that exhibited insertion polymorphisms and substantial copy number variation between strains, an absence of target site duplications (TSDs) upon insertion, and split terminal repeats that were unlike any other described TE at the time. Ferris (1989) described *Gulliver*, a 12.2 kb DNA transposon characterised by 15 bp terminal inverted repeats (TIRs) and 8 bp TSDs, which exhibited insertion polymorphisms between laboratory strains and complete absence in certain field isolates. *TOC1* and *Gulliver* have frequently been used as experimental models of transposition in *C. reinhardtii*, for example in studies exploring gene silencing mechanisms (Wu-Scharf et al. 2000; Jeong et al. 2002; Casas-Mollano et al. 2008). A handful of other TEs followed: the DNA transposons *Tcr1* (Schnell and Lefebvre 1993; Ferris et al. 1996), *TOC2* (Day 1995) and *Tcr3* (Wang et al. 1998), and a second unusual retrotransposon *Pioneer1* (Graham et al. 1995). Later experimental work characterised the *Gypsy* long terminal repeat (LTR) element *REM1* (Perez-Alegre et al. 2005), the non-autonomous DNA transposon *Bill* and the non-autonomous retrotransposon *MRC1* (Kim et al. 2006).

The availability of preliminary *C. reinhardtii* assembly versions (Grossman et al. 2003) enabled researchers to directly identify and curate repetitive sequences of interest. Although nonautonomous, *TOC1* was found to be related to other TEs (e.g. *TOC3*) that contained similar split terminal repeats and encoded proteins with RT and tyrosine recombinase (YR) domains, placing these elements within the emerging *Dictyostelium* intermediate repeat sequence-like (DIRS) group (Goodwin and Poulter 2004). Although it lacks split terminal repeats, *Pioneer1* was also classified as a DIRS element based on the identification of a YR domain (Goodwin and Poulter 2001; Poulter and Goodwin 2005). Kojima and Fujiwara (2005) described several families of long interspersed nuclear elements (LINEs) forming the new clade *Dualen*, which unlike all other LINEs encode both restriction-like endonuclease (RLE) and apurinic/apyrimidinic endonuclease-like endonuclease (APE). Cognat et al. (2008) annotated short interspersed nuclear elements (SINEs), most of which rely on *Dualen* elements for their activity. The most thorough annotation effort was performed by Kapitonov and Jurka as part of the v3 genome assembly (Merchant et al. 2007), with their consensus sequences forming the vast majority of the 119 *C. reinhardtii* TEs deposited in Repbase (https://www.girinst.org/repbase/). This work resulted in the description of *Novosib*, a new superfamily of DNA transposons (Kapitonov and Jurka 2008; Yuan and Wessler 2011). Annotations of other important TEs were also improved, for example *Gulliver* was classified as an autonomous member of the *hAT* superfamily based on the identification of its transposase (Kapitonov and Jurka 2006a). The v4 and v5 assemblies have continued to be sources of biological novelty, most recently demonstrated by the contribution of *C. reinhardtii* sequences to the discovery of the *Helitron2* group (Bao and Jurka 2013) and the *Kyakuja-Dileera-Zisupton (KDZ)* superfamily of DNA transposons (Böhne et al. 2012; Iyer et al. 2014).

Unfortunately, there has been very little attention paid to the genome-wide distribution of TEs in *C. reinhardtii*, and no systematic study of TEs from a population perspective.

Philippsen et al. (2016) reported an underrepresentation of TEs in *C. reinhardtii* introns, suggesting that the long intron lengths found in the species (1.3.2) cannot be attributed to recent invasion by TEs. They also found that introns located near the start of genes contained fewer TE insertions than those in central or 3' locations. In Chapter 5, I present a major update to TE annotation in *C. reinhardtii* and describe a new clade of *Penelope*-like elements (PLEs) discovered in the species. I summarise the genomic distribution of TEs and other repetitive sequences in Chapter 4.

1.3.8 Structural variation between laboratory strains and field isolates

Structural variation (SV) is often defined as variants >50 bp, and includes large indels, inversions, duplications, copy number variants (CNVs) and transpositions. Both Flowers et al. (2015) and Gallaher et al. (2015) performed brief analyses of SV using their re-sequencing data. Considering field isolates, Flowers et al. (2015) estimated that on average field isolate genomes contained 32 genes encoding proteins with recognisable domains that are absent from the reference genome (i.e. presence-absence variants, PAVs). Many of these PAV genes were from large gene families in *C. reinhardtii*, such as the scavenger receptor cysteine-rich (SRCR) and C-type lectin (CTL) domain families (Wheeler et al. 2008). Using a coverage-based analysis, they also identified several large regions exhibiting CNV. Considering SV between laboratory strains, it is expected that differences between the two different haplotypes (1.2.3) likely represent variants segregating in the population, while differences between copies of the same ancestral haplotype represent mutation in the laboratory. Gallaher et al. (2015) identified more than 4,000 putative SVs present between alternate haplotypes, ~16% of which were predicted to disrupt genes, and a further 800 between copies of the same haplotype. They also matched the sequences of eight TEs shown experimentally to be active (1.3.7) to deletions identified in strains relative to the reference, calling 84 TE polymorphisms in total. Flowers et al. (2015) identified a small number of regions that appear to have undergone duplication in the laboratory, for example an ~400 kb segment of chromosome 1 in CC-407. I curate structural mutations that have occurred in the laboratory via a direct comparison between long read genome assemblies in Chapter 4.

1.3.9 Organelle genomes

Briefly, the *C. reinhardtii* plastid genome is circular, 205.5 kb in length, with a GC content of 34.6% (Maul et al. 2002; Smith and Lee 2009; Gallaher et al. 2018). It is estimated to be present in ~80-90 copies per cell (Misumi et al. 1999; Gallaher et al. 2018). The overall mutation rate in the plastid genome has been estimated as $\mu = 9.23 \times 10^{-10}$ (Ness et al. 2016), which is not significantly different from the nuclear genome (1.3.4). The plastid genome is uniparentally inherited from the MT^{+} parent, although there is evidence for recombination or "leaky" inheritance between the mating types (Boynton et al. 1987; Dürrenberger et al. 1996; Hasan et al. 2019).

The mitochondrial genome is linear, 15.8 kb in length, has a GC content of 45.2% and is present in >100 copies per cell (Vahrenholz et al. 1993; Gallaher et al. 2018). The mitochondrial genome is inherited from the *MT*$^{-}$ parent, and unlike the plastid, inheritance appears to be strictly uniparental (Nakamura 2010; Hasan et al. 2019). Although beyond the scope of this introduction, it is interesting to note that both linear and circular mitochondrial genomes are observed in the volvocine algae and there appear to have been several evolutionary transitions between the two architectures (Hamaji et al. 2017; Smith and Craig 2020).

1.3.10 Genome assemblies of related species

Algal genomics is a rapidly growing field (Blaby-Haas and Merchant 2019). There were just six chlorophyte genome assemblies uploaded to GenBank by the end of 2010, a number that had risen to 22 by the end of 2015, and 131 by the end of 2020. These assemblies range from highly contiguous reference quality assemblies (e.g. Worden et al. (2009), Blanc et al. (2012), Roth et al. (2017)) to highly fragmented draft assemblies (~60% of chlorophyte assemblies in GenBank have a contig-level N50 <50 kb). Recent studies have collectively sequenced over 100 genomes at once (Nelson et al. 2019; Nelson et al. 2020) and the number of algal genomes is expected to continue to rise substantially. This growth has led to the establishment of PhycoCosm, a dedicated algal genome browser (Grigoriev et al. 2021).

Considering species more closely related to *C. reinhardtii*, sequencing effort has largely focussed on the multicellular volvocine algae, with recent years seeing assemblies for *V. carteri* (Prochnik et al. 2010), *G. pectorale* (Hanschen et al. 2016), *T. socialis* (Featherston et al. 2018), and *Y. unicocca* and *Eudorina* sp. (Hamaji et al. 2018). Draft assemblies for two unicellular core-*Reinhardtinia* (1.2.2) species, *E. debaryana* and *C. sphaeroides*, were produced by Hirashima et al. (2016), while a third assembly for an undescribed *Chlamydomonas* species was included in the large dataset of Nelson et al. (2019). These assemblies have ranged between ~110 Mb and 185 Mb, with genome-wide average GC contents of 56% to 68%. The assemblies that have been annotated are also all intron-rich, suggesting certain features of genome architecture are likely to be widely conserved across the clade. In Chapter 3, I produce the first genome assemblies for close relatives of *C. reinhardtii* and characterise genome architecture and gene content in *Chlamydomonas* (*sensu stricto*) and the volvocine algae.

1.4 Introduction to Core Methodology

1.4.1 Whole-genome re-sequencing and variant calling

Our ability to sequence nucleic acids has been revolutionised over the past two decades. The technological advances that have resulted in higher accuracy and throughput, and crucially plummeting costs, have transformed several fields of biological research. Writing in 2006, Bentley estimated the cost of sequencing a human genome to be $10 million, while the original Human Genome Project completed three years earlier was estimated to have cost $1 billion. We are now at or approaching the $1,000 human genome (National Human Genome Research Institute 2020), although this figure solely reflects the price of sequencing and excludes several associated costs (Schwarze et al. 2020). This staggering reduction in cost in the last 15 years has been driven by the transition from classical Sanger sequencing (Sanger et al. 1977b) to next-generation sequencing (NGS), reviewed by Goodwin et al. (2016) and Levy and Myers (2016). NGS is currently dominated by Illumina sequencing (Bentley et al. 2008), which offers read lengths of 50-300 bp (most commonly 100-150 bp) with average error rates of substantially less than 1/1,000 bp. Reads can be single or paired-end, where paired-end refers to the sequencing of both flanks of a single DNA fragment (generally <500 bp, although fragments of several kb can be sequenced in "mate pair" libraries). There is a myriad of potential NGS applications, including whole-genome sequencing for *de novo* assembly, transcriptome sequencing (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq), bisulfite sequencing and exome sequencing. When NGS is performed on genomic DNA with a view to comparing the sequenced sample to an existing reference genome, it is referred to as whole-genome re-sequencing.

The advances of NGS have driven evolutionary biology, and particularly population genetics, to develop from a largely theory-based discipline to a field in which enormous data sets are regularly produced and analysed (Pool et al. 2010). Prior to NGS, obtaining estimates of genetic diversity and other associated metrics either required Sanger sequencing of multiple loci from multiple individuals (or rarely whole genomes e.g. Begun et al. (2007)) or developing SNP arrays for genotyping, assuming the availability of a reference genome. In other fields such as molecular ecology, reference-free approaches including microsatellite analysis and AFLP were, and in some cases still are, widely used (Schlötterer and Pemberton 1998; Bensch and Akesson 2005). At an appropriate coverage (see below), whole-genome re-sequencing enables NGS reads to be mapped to a reference genome, and subsequently variants (generally SNPs and small indels) to be called between the reference and the sequenced sample. These variants can then be used as the basis for performing population genetics analyses. Although not discussed herein, there is considerable flexibility in this process, for example with lower coverage datasets researchers can work with genotype likelihoods rather than variant calls (Korneliussen et al. 2014), and reference-free NGS approaches such as restriction site-associated DNA sequencing (RADseq) have been developed (Andrews et al. 2016).

The process of whole-genome re-sequencing and variant calling can be summarised in five major steps. First, an appropriate re-sequencing strategy must be selected and performed. The desired coverage will depend on the ploidy and heterozygosity of the sample in question, as at lower coverages there can be substantial sampling variation in the proportion of reads obtained from each haplotype. Kishikawa et al. (2019) estimated that with contemporary sequencing and variant calling approaches, an average coverage of ~15x was appropriate for SNP calling in humans, although depths several times higher were required to call indels accurately. Assuming equal error rates, paired-end sequencing and longer read lengths are both expected to improve read mapping due to the additional information they provide. NGS sequencing is also biased in regions of extreme base composition (Dohm et al. 2008; Hillier et al. 2008), and in species with high GC or AT contents it is important to consider adjusting library preparation, for example by amending PCR conditions or removing the PCR step entirely (Aird et al. 2011). As a second step, sequencing reads should be checked for general quality metrics, which can be performed using tools such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The raw reads can be pre-processed to remove low quality bases or sequencing adapters, although the importance of read trimming on variant calling has been questioned (Bush 2020).

Third, reads are mapped to the reference assembly, which is achieved using read aligners such as Bowtie 2 (Langmead and Salzberg 2012) and most commonly BWA-MEM (Li and Durbin 2009; Li 2013). Read alignments are stored in the Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM) format (Li et al. 2009), which provides rich information for each individual read alignment, including the genomic coordinates and strand of mapping, mismatches between the read and reference, the mapping quality, whether the read maps uniquely, whether the mate of the read is aligned as a proper pair, and if the read is "clipped" (i.e. only part of the read aligns to a given region and bases from the 5' and 3' of the read are not part of the alignment). Samtools (Li et al. 2009) is a well-developed bioinformatics toolkit for manipulating SAM/BAM files (e.g. sorting and merging files, extracting specific regions, calculating alignment depth, etc.). Additional post-alignment processing steps are recommended by certain pipelines, one of the most common being the identification of "duplicate" reads that are assumed to be derived from the same DNA fragment and therefore represent non-independent observations.

Fourth and fifth, the information contained within the read alignments is converted to variant calls, which are subsequently filtered based on assessments of their quality. Several pipelines are commonly used for variant calling, including the Genome Analysis Toolkit (GATK) (McKenna et al. 2010; DePristo et al. 2011), freebayes (Garrison and Marth 2012) and samtools itself. Pre-variant calling procedures are usually implemented to avoid biases from alignment artefacts, especially around indels. For example, GATK performs local re-alignment around indels and recalibrates base quality scores, while samtools calculates a base quality score (BAQ) that relates to the probability that the base is misaligned. More recently, tools such as GATK HaplotypeCaller (Poplin et al. 2018) perform local re-assembly around variant regions, subsequently calling variants from inferred haplotypes rather than directly from the read alignments. Variant calling is improved by analysing multiple samples collectively, since population-level information on variants can be incorporated e.g. a low coverage variant detected in a given sample is more likely to be genuine if that variant has been called as high-quality in other samples. The resulting variant and invariant sites are represented in a Variant Call Format (VCF) file (Danecek et al. 2011), and each variant has several associated metrics, which in GATK includes the number of reads supporting the reference and alternative alleles, quality scores (both for the entire site and individual genotype calls) and information on read mapping biases that may introduce false positives (e.g. strand and read position biases). While it is possible to perform variant filtering relative to a set of highly trusted variants for certain model organisms, for most species variant filtering is performed *ad hoc* based on the variant quality metrics. This often requires careful data exploration to assess the distribution of variant quality metrics in the sample in question and to avoid introducing biases. Crucially, it is imperative that consistent filtering criteria are applied to both variant and invariant sites.

Finally, although such analyses are not featured in this thesis, it is important to note that several pipelines have been developed to call SVs from NGS data (Kosugi et al. 2019; Mahmoud et al. 2019). These approaches often use the information provided by discordant read pairs and clipped reads, although due to the inherent complexity of SVs both false positive and negative rates are generally high and can approach 90% for certain types of SVs (Mahmoud et al. 2019). Long read sequencing technologies are expected to substantially improve SV calling (1.4.2). Another development is the production of graph-based pangenomes, which incorporate several genomes from the same species and remove much of the reference bias in read mapping and variant calling (Eizenga et al. 2020).

In Chapter 2, I use whole-genome re-sequencing and variant calling to identify SNPs and subsequently characterise patterns of population structure amongst field isolates of *C. reinhardtii*. I also use estimates of genetic diversity in Chapters 3 and 4 to assess the coding potential of genes.

1.4.2 Genome assembly using long read sequencing

As with whole-genome re-sequencing (1.4.1), NGS has transformed genome sequencing and assembly. Early genome assemblies were based on Sanger sequencing, including the first genome assembly of a bacteriophage (Sanger et al. 1977a), the first multicellular eukaryote genome in *C. elegans* (*C. elegans* Sequencing Consortium 1998), and indeed the *C. reinhardtii* genome itself (1.3.1). Sanger-based genome projects were large-scale, expensive and time consuming, although for several reasons the assemblies they produced were generally of high quality (and in some cases complete e.g. yeast (Mewes et al. 1997)). Furthermore, since Sanger-based assemblies were generally produced for well-studied organisms, in many cases the assembled scaffolds could be placed on chromosomes by incorporating extrinsic evidence from linkage maps. In contrast, the low-cost and high-throughput of NGS has opened genome sequencing to a wide range of non-model organisms (1.3.10). However, NGS-based assemblies are often highly fragmented, since it is impossible to assemble across repetitive sequences with short read lengths.

Recently, new approaches have been developed that sequence single molecules of DNA. Often called third generation sequencing, both the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT, also “Nanopore”) platforms are capable of sequencing reads tens of kb in length (van Dijk et al. 2018), and in certain instances Nanopore sequencing can sequence ultra-long reads hundreds of kb in length (Jain et al. 2018b). However, these technologies have far higher error rates (5-15%) than NGS (Watson and Warr 2019). Nonetheless, since the read-lengths are longer than many of the repetitive sequences present in genomes, the assembly contiguity that can be achieved using long read sequencing is often several times higher than that of NGS-based assemblies. Such improvements have many functional implications beyond simply assembling repetitive sequence, for example many genes can be fragmented or entirely unassembled due to gaps in NGS assemblies (Thomma et al. 2016). Rapid technological advances in third generation sequencing continue to result in increased throughput, longer read lengths and reduced error rates, and it is expected that long reads will form the basis of high-quality genome assemblies across an ever-increasing array of organisms in the coming years. Most recently, PacBio have developed HiFi sequencing, which is capable of sequencing reads >20 kb with error rates <1%. HiFi reads have been used to produce an inbred mouse genome with an N50 >20 Mb, and even assemble a >35 Gb assembly of the giant redwood with an N50 >5 Mb (Cheng et al. 2020). We are now entering an era where complete telomere-to-telomere chromosomal assemblies are possible, even for large and moderately repetitive genomes such as human (Miga et al. 2020).

A number of long read assemblers have been developed, including Canu (Koren et al. 2017) and wtdbg2 (Ruan and Li 2020), some of which are haplotype aware and attempt to produce phased assemblies e.g. FALCON-Unzip (Chin et al. 2016). Relatively high coverage is generally required to achieve contiguous assemblies, for example Canu assemblies continue to improve up to ~50x coverage (Koren et al. 2017). Most long read assemblers follow the "overlap-layout-consensus" paradigm. Briefly, the "overlap" stage involves comparing reads and identifying overlaps, the "layout" stage identifies contiguous stretches of overlapping reads to form contigs, and the "consensus" step attempts to call the most likely sequence given the underlying errors and variation in the reads. This approach differs from NGS assemblers that are generally based on de Bruijn graphs (Compeau et al. 2011). Although the resulting assemblies can be highly contiguous, the consensus sequence has a high error rate (especially for indels) and requires additional "polishing" (Watson and Warr 2019). Error correction is generally performed first by mapping the raw long reads to the *de novo* assembly to improve consensus calls, which for PacBio assemblies is performed by the Genomic Consensus module (https://github.com/PacificBiosciences/GenomicConsensus). Second, NGS data can be mapped to the polished assembly and additional error correction performed using tools such as Pilon (Walker et al. 2014). Both polishing steps can be iterated, although repeated polishing may reduce assembly quality (Miller et al. 2018). The final polished assembly can be scaffolded using any available linkage data or extrinsic evidence. Although not used in this thesis, new "linked read" technologies (e.g. 10X and Hi-C) provide high-throughput alternatives to genetic mapping for obtaining linkage information (van Dijk et al. 2018).

Finally, it is important to note that long reads are not only used for *de novo* assembly. As single molecules of DNA are sequenced without amplification, it is possible to detect epigenetic base modifications from the raw sequence data (Flusberg et al. 2010; Liu et al. 2019). Long read sequencing can be performed on cDNA (PacBio and Nanopore) or directly on RNA (Nanopore) to sequence full-length transcripts. This can greatly aid gene annotation and has been used to reveal non-canonical features such as polycistronic genes (Gordon et al. 2015; Gallaher et al. 2021). As already introduced (1.4.1), long reads also considerably outperform NGS data in calling SVs (Mahmoud et al. 2019). With the onset of highly accurate long read sequencing methods such as HiFi, we can also expect to see long reads used to call SNPs and small indels, since accurate long reads will enable variants to be detected in repetitive or "dark" genomic regions where NGS data does not successfully map.

In Chapter 3, I use PacBio sequencing to produce highly contiguous genome assemblies for three unicellular relatives of *C. reinhardtii*, producing genomic resources that enable comparative genomics analyses. In Chapter 4, I use PacBio sequencing to update the *C. reinhardtii* reference genome, producing near-complete assemblies for two laboratory strains. In Chapter 5, I produce a PacBio assembly for a field isolate of *C. reinhardtii,* which I use to identify and annotate polymorphic TEs.

1.4.3 Transposable element classification and annotation

TEs are selfish mobile units of DNA that are present in all domains of life and are a near universal constituent of eukaryotic genomes. TEs are evolutionary ancient and have originated several times independently, displaying an outstanding diversity of form and mechanism. Given their ubiquity, TEs play a fundamental role in genome evolution and have been implicated in numerous evolutionary phenomena. Via their ability to replicate and increase in copy number, TEs can be a driver of genome size evolution (Lynch 2007). TE insertions are a major source of SV both within and between species, and both active and inactive TEs are associated with genomic rearrangements (Bourque et al. 2018). The co-option of TE sequence has been demonstrated to be a pervasive feature of regulatory sequence evolution (Chuong et al. 2017) and numerous examples of TE proteins undergoing host domestication have been documented (Jangam et al. 2017). TEs have also been hypothesised to play a role in hybrid incompatibility and speciation (Jurka et al. 2011). More generally, knowledge of the TE sequence in a genome can be an invaluable resource for exploring features of genome architecture. For example, the identification of regions rich in TEs (or particular types of TEs) can be indicative of heterochromatin (Lippman et al. 2004) and major genomic features such as centromeres (Wong and Choo 2004). Finally, repeat masking a genome is a prerequisite of gene annotation and therefore an integral part of genome projects. In order to study TEs and their wider biological interactions, TE annotations for the species of interest (or a closely related species) are required. As described in detail below, this generally involves the production of a library of TE consensus sequences, with each sequence ideally classified by TE type.

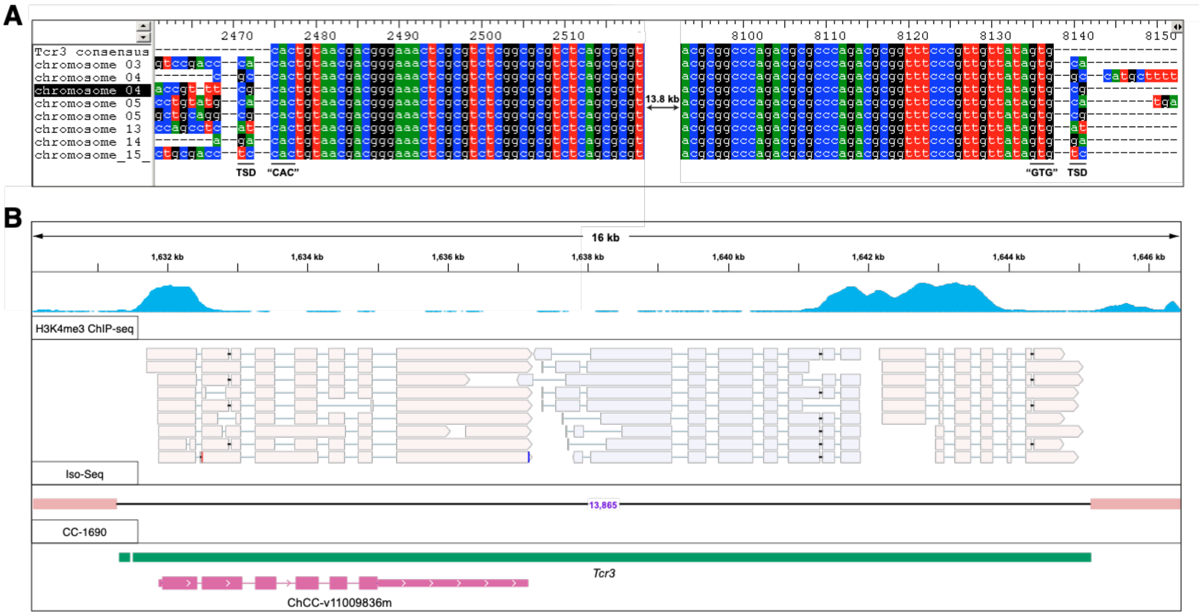
As a result of their diversity, several TE classification systems have been proposed (Wicker et al. 2007; Kapitonov and Jurka 2008; Arkhipova 2017), which generally aim to group TEs based on their phylogenetic relationships and shared mechanisms. The highest rank in traditional classifications is class, which divides TEs based on their transposition intermediate, RNA (class I, also retrotransposons) or DNA (class II) (Finnegan 1989). Furthermore, most types of TEs can be either autonomous, which encode the proteins necessary for their transposition, or non-autonomous, which utilise proteins produced by independent autonomous TEs (McClintock 1956). Retrotransposons transpose via a copy-and-paste mechanism and all autonomous retrotransposons encode proteins containing an RT domain. Although the RT domain is monophyletic (Arkhipova 2017), retrotransposons can be divided into four evolutionary ancient and distinct groups (sometimes called orders or subclasses), namely LINEs, LTRs, DIRS and PLEs, each of which is clearly differentiated by the presence of unique additional protein domains, structural motifs and mechanisms of transposition. A fifth group, SINEs, are strictly non-autonomous and rely on the activity of LINEs (Kramerov and Vassetzky 2011). Additionally, a new group of non-autonomous retrotransposons termed retrozymes has been described, which may be grouped with SINEs depending on interpretation (de la Peña et al. 2020). Retrozymes rely on the machinery of autonomous LTR elements in plants (Cervera et al. 2016) and most likely PLEs in animals (Cervera and de la Peña 2020). Class II elements originally referred specifically to DNA transposons, an extremely diverse TE group that transpose via a cut-and-paste mechanism facilitated by a transposase enzyme. However, as genomes of more diverse species have been assembled and annotated, three additional independent class II groups have emerged, the Helitrons (Kapitonov and Jurka 2001), Maverick/Polintons (Feschotte and Pritham 2005; Kapitonov and Jurka 2006b) and Cryptons (Goodwin et al. 2003).

All major groups of both class I and II TEs can be further divided into ancient major clades, often termed superfamilies. Many superfamilies exhibit distinguishing features that can be used for classification. For example, there are more than 20 described superfamilies of DNA transposons (Kojima 2020), many of which can be distinguished based on the lengths of TSDs, the length and sequence of TIRs, and insertion biases. *Mariner* elements have "YR..YR" in their TIR termini, insert at "TA" sites and produce 2 bp "TA" TSDs; *hAT* elements have "YA..TR" termini and produce TSDs of 5-8 bp; *EnSpm* elements have "CAC..GTG" termini and produce TSDs of 2-4 bp, and so on. Certain retrotransposon superfamilies can be distinguished by the presence or order of protein domains. As introduced in 1.3.7, *Dualen* LINE elements encode proteins with two endonucleases, unlike any other LINEs. *Copia* LTR elements can be distinguished from all other LTRs as the integrase domain is located upstream of RT rather than at the C-terminus of the protein. However, not all superfamilies can be so clearly distinguished, and in many cases phylogenetic analyses of protein domains may be required. The online tool RTclass1 (Kapitonov et al. 2009) automates this process for LINEs. Such an approach is obviously only applicable to autonomous TEs, and the classification of nonautonomous TEs can often be challenging as a result.

Within superfamilies, multiple copies of very closely related TEs can be grouped as a family or subfamily. Each family can be represented as a single consensus sequence, which aims to re-create the ancestral sequence of a TE when it was actively transposing and creating new copies in a given genome or genomes. There is no clear definition of what constitutes a TE family, although the "80-80-80 rule" is often applied (Wicker et al. 2007). This rule states that two TEs belong to the same family if they exhibit at least 80% sequence similarity over at least 80% of the internal region, termini (to account for solo LTRs, 5' truncation, etc.), or both, and finally are at least 80 bp in length. The subfamily classification only becomes relevant if there are two or more subpopulations of TEs within the same family. To give an example, if there were two active TE populations in a genome that are divergent by 10% on average between the populations and far less than 10% within the populations, then these would represent a single family under the 80-80-80 rule, but distinct subfamilies.

Family or subfamily TE consensus sequences form the basis of TE libraries. Consensus sequences can be produced by several means, which can largely be considered as either automated or manual. One of the most common automated approaches, implemented in the widely used RepeatModeler (Smit and Hubley 2008-2015), is to identify and cluster repetitive sequences based on their occurrence in the input genome. However, due to the complexity of TEs this approach has several shortfalls. First, the consensus families are often fragmented, for example due to 5' truncation in LINEs or the presence of solo LTRs (Flynn et al. 2020). Second, low copy number TEs cannot be detected. Third, there is no requirement that the repeats are TEs and large gene families such as histones can be included. Fourth, the resulting sequences still need to be classified, which is usually based on homology to known curated TEs. This introduces a substantial phylogenetic bias, and classification is expected to be far less reliable in species that are very distantly related to any species for which there are curated TE libraries. An alternative automated approach is to identify individual copies of TEs based on structural motifs, which can later be clustered and represented as consensus sequences if desired. Examples of this approach include LTRharvest (Ellinghaus et al. 2008), LTR_retriever (Ou and Jiang 2018) and HelitronScanner (Xiong et al. 2014). While this approach has the advantage of capturing more complete TE sequences, there are once again several limitations. First, not all TEs contain structural motifs that can be modelled. Second, degraded TE copies may not be identified. Third, it is only possible to model what is known, which once again introduces circularity with respect to manual curation. For example, it is not clear that the recently described *Helitron2* superfamily present in *C. reinhardtii* (1.3.7) can be identified by HelitronScanner. Recently, RepeatModeler was extended to incorporate structural identification of LTRs (Flynn et al. 2020), and a combined approach is likely to yield the best results if pursuing automation. Manual curation remains the gold standard for TE annotation, especially in species other than vertebrates and angiosperms. Curated consensus sequences are available from repositories including Repbase and Dfam (https://dfam.org/).

To manually curate TEs, one must first obtain TE candidates. This can be achieved by running a tool such as RepeatModeler. Alternatively, if the genome is already annotated it may be possible to identify candidate genes that encode proteins with putative TE domains. If a genome assembly for a different individual of the same species (or a very closely related species) is available, it is also possible to treat large indels identified between the assemblies as potential TEs. Similarly, if population re-sequencing data is available it is possible to call deletions relative to the reference (which are identified far more accurately than other SVs with NGS data), which can be treated as potential TE insertions in the reference (or excisions in the NGS sample) (Uzunović et al. 2019). A dataset of candidates can then be queried against the target genome using megablast (Camacho et al. 2009) in order to retrieve multiple copies. As the input sequences may be truncated, the coordinates of hits can be extended at the flanks by several kb, before being aligned. The resulting alignment is then viewed in a sequence editor and curated. If the sequence appears to be a TE and the sequence is complete (the flanks can be extended if not), then after removing any poorly aligned sequences a consensus sequence can be produced using a number of available tools. Finally, the consensus sequence should be checked relative to the original alignment to ensure its accuracy. TEs can then be classified as presented above. This approach approximately follows that performed by Suh et al. (2014).



```
                      2470      2480      2490      2500      2510              8100      8110      8120      8130      8140      8150
Tcr3 consensus  ----------------cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg-----------------
chromosome 03   gtccgacc--ca---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--ca-------------
chromosome 04   -------c--gc---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--gc--catgctttt
chromosome 04   accgt-tt--cg---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--cg-------------
chromosome 05   cctgtatg--ca---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--ca---------tga
chromosome 05   gctgcagg--cg---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--cg-------------
chromosome 13   ccagcctc--at---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--at-------------
chromosome 14   -------a--ga---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--ga-------------
chromosome_15_  ctgcgacc--tc---cactgtaacgacgggaaactcgcgtctcggcgcgtctcagcgcgt  acgcggcccagacgcgcccagacgcggtttcccgttgttatagtg--tc-------------
```

Figure 1. Producing a consensus sequence for the active DNA transposon *Tcr3*.
(A) Bioedit (https://bioedit.software.informer.com) screenshot of the consensus sequence (top row) and the alignment of individual copies (subsequent rows). Only the left and right termini of the alignment are shown. TSDs and the termini motifs are highlighted.
(B) IGV view of a single polymorphic copy of *Tcr3* on chromosome 4 of the reference genome (this copy is highlighted by a black box in panel A). The top track shows H3K4me3 ChIP-seq data, followed by Iso-Seq data, alignment of the CC-1690 genome, and gene and repeat annotations. Note that the Iso-Seq reads are all multimappers (since the TE is present in many copies) and it cannot be confirmed that this particular copy is actively expressed.

To provide an example, Figure 1A shows the left and right flanks of the *C. reinhardtii* DNA transposon *Tcr3*, which is active in laboratory strains (Wang et al. 1998). The consensus sequence is shown on the top row of the alignment, and the aligned TE copies extracted from the genome and used to produce the consensus are shown below. The sequence beyond the flanks is unalignable, which is indicative or a repetitive element, although a 2 bp TSD can be observed. The TIRs terminate in the "CAC..GTG" motif, which together with the 2 bp TSDs suggests that *Tcr3* is a member of the *EnSpm* superfamily (see above). As *Tcr3* is active, it is also possible to check individual copies for evidence of polymorphism and expression using genome browsers such as the IGV (Robinson et al. 2011). Figure 1B shows a single copy of *Tcr3* that is present on chromosome 4 of the reference genome and is polymorphic between the reference and the CC-1690 genome (O'Donnell et al. 2020). Although in this case the termini of *Tcr3* were already obvious from the alignment (Figure 1A), in many cases termini may be less defined and polymorphisms can provide useful confirmatory evidence. *Tcr3* also appears to contain three expressed genes, as shown by the H3K4me3 ChIP-seq peaks (which mark active promoters in *C. reinhardtii* (Ngan et al. 2015)) and full-length PacBio sequenced cDNAs (i.e. "Iso-Seq"). Based on the Iso-Seq reads it is possible to determine open reading frames (ORFs) for each gene, and in this case one of the genes was included in the gene annotation itself. The predicted proteins can then be checked against known TE proteins using homology searches (e.g. blastp) or phylogenetic analysis if required. As expected, the first gene in *Tcr3* encodes a transposase with homology to known *EnSpm* transposases, confirming classification. Although for many species the functional data used in Figure 1B is unavailable, the initial alignment of copies is sufficient in most cases to produce a consensus sequence and perform classification.

In Chapter 5, I perform an exhaustive annotation of TEs in *C. reinhardtii*, significantly extending the exiting annotations (1.3.7). In Chapter 3, I perform targeted annotation of the most abundant TEs in the *C. incerta*, *C. schloesseri* and *E. debaryana*, which I use to improve gene annotation and to characterise putatively centromeric sequences.

1.5 *Chlamydomonas* in a wider evolutionary context

As presented herein, *C. reinhardtii* is an established and important model system in plant biology and several other specific areas. However, very little is known about the evolutionary biology of the species in general and several key genomics resources are lacking relative to more established model systems such as *Drosophila melanogaster* and *Arabidopsis thaliana*. There are two primary aims of the work presented in this thesis. First, I aim to address several questions relating to the evolutionary genetics and genomics of *C. reinhardtii*, some of which have wider relevance to other taxa. Second, I aim to produce the necessary genetic and genomic resources to enable *C. reinhardtii* and *Chlamydomonas* to be further developed as a study system in evolutionary research. In a wider context, I aim to bring *C. reinhardtii* more closely in line to other model species with respect to these objectives. In this section, I outline the questions addressed and resources produced in this thesis, before briefly introducing some of the most interesting evolutionary questions that could be addressed using *C. reinhardtii* in the future. The work presented in this thesis is more substantially contextualised in the introduction sections of each subsequent results chapter (2-5) and is discussed in a broad sense in Chapter 6.

As introduced in 1.2.4, there is currently no suitable population genetics dataset for *C. reinhardtii*, with at most four isolates sequenced from one location and population structure evident between isolates from different locations. This is in stark contrast to many other model systems, in which tens or even hundreds of individuals have been subject to whole-genome re-sequencing. In Chapter 2, I address this deficit by introducing whole-genome re-sequencing data for more than 20 isolates sampled from a single site in Quebec. I explore how genetic diversity is geographically structured among the currently available field isolates, drawing inferences on the demography and ecology of the species. This is important in a wider context since there are very few population genomics datasets available from microbial eukaryotes, and there has been longstanding debate concerning the nature of population structure and demographic processes in such species.

Another shortfall relative to most model systems is the complete lack of genomic resources for any close relatives of *C. reinhardtii*. This has contributed to a relatively poor understanding of the phylogenetic relationships among related species (1.2.2) and of genome architecture and evolution in general (1.3.2). The availability of outgroup genomes is central to several key analyses in molecular evolution, and comparative genomics is a powerful tool for identifying functional sequence. In Chapter 3, I produce genome assemblies and annotations for close relatives of *C. reinhardtii* with a view to enabling comparative genomics analyses to be performed. I use these resources to characterise several features of the *C. reinhardtii* genome within an evolutionary framework, including centromeres, false positive and previously uncharacterised gene annotations, and conserved noncoding elements.

Despite being among the first eukaryotic genomes to be sequenced, the *C. reinhardtii* assembly and annotations are not of comparable quality to those of most model organisms, and they have not been updated using long read sequencing approaches. The quality of the genome assembly and accuracy of the gene model annotations underly most contemporary research using *C. reinhardtii*, and indeed green algae in general, and their improvement would be expected to have impact far beyond evolutionary biology. In Chapter 4, I assemble near-complete genome sequences for both MT^{+} and MT^{-} laboratory strains and introduce improved gene annotations. These assemblies reveal several errors in past genome versions that would be expected to confound analyses requiring the accurate assembly of chromosomes.

Finally, the genomics of chlorophytes, and indeed microbial eukaryotes in general, are severely understudied relative to most animal and plant lineages, and many fungi. Thus, one of the most exciting aspects of developing *C. reinhardtii* as an evolutionary model is the potential discovery of novel genetic and genomic features and phenomena that have been overlooked in other taxa. The *C. reinhardtii* genome is known to harbour a substantial diversity of TEs (1.3.7), many of which were first discovered in the species. In Chapter 5, I perform exhaustive curation of TEs in the species, substantially expanding upon previous annotations. I discover and describe a major new clade of PLEs that are present in other green algae, plants, protists and animals. This finding has several implications for genome evolution in these taxa, and substantially elevates the evolutionary importance of PLEs relative to more widely studied TEs.

Considering future research, many of the most promising applications of *C. reinhardtii* stem from its experimental amenability, particularly in combining experimental and sequencing approaches. As described in 1.2.5, such approaches have so far been used to characterise the mutation rate and spectra of SNMs and small indels, to estimate the distribution of fitness effects of these mutations, and to study GC-biased gene conversion. There is substantial potential to investigate recombination and related phenomena using experimental designs involving the re-sequencing of progeny produced from highly replicated crosses. The high genetic diversity between strains provides high marker density relative to the read lengths of standard NGS (~2-3 SNPs per 100 bp read) and the haploid state simplifies analysis and negates the need for haplotype phasing. At least two projects are currently underway to characterise fine-scale variation in recombination rate, with possible applications including the identification of genomic features that correspond to rate variation and testing for a mutagenic effect of recombination. Furthermore, combining empirical data of mutation, selection and recombination with population genetics datasets may present opportunities to better understand the interplay of evolutionary forces that influence genome-wide variation in genetic diversity. Similarly, the MA lines generated for *C. reinhardtii* present a rare opportunity to quantify the rate of transposition for several different types of TEs. This is currently being approached using both short and long read sequencing technologies. It would also be possible to quantify transposition rates in multiple environmental conditions (e.g. temperature, salinity) via additional MA experiments. There is ample, although sometimes conflicting, evidence of the effect of environmental conditions on transposition in *D. melanogaster* (Guerreiro 2012) and *A. thaliana* (Quadrana et al. 2016), although to my knowledge this has never been tested under MA conditions. Although these examples demonstrate that *C. reinhardtii* has potential for valuable evolutionary research in specific areas, to fully interpret, contextualise and develop the results yielded from such experiments we require a far greater general understanding of the evolutionary genetics and genomics of the species and its closest relatives. The work in this thesis takes some of the first steps towards achieving this goal.

Chapter 2

Patterns of Population Structure and Complex Haplotype Sharing Among Field Isolates of the Green Alga *Chlamydomonas reinhardtii*

2.1 Preface

The work in this chapter has been published as a manuscript in *Molecular Ecology* and the first-person plural is used throughout to maintain consistency. Minor changes have been made to the published version to preserve formatting across the thesis. I performed all analyses, wrote the first draft of the manuscript, and produced all figures and tables. Katharina Böndel performed preliminary analyses and discovered the presence of the identical by descent tracts. Takashi Nakada, Takuro Ito and Graham Bell performed isolate sampling. Rob Ness, Kazuharu Arakawa and I performed DNA extraction and sequencing.

Citation:

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.

2.2 Abstract

The nature of population structure in microbial eukaryotes has long been debated. Competing models have argued that microbial species are either ubiquitous, with high dispersal and low rates of speciation, or that for many species gene flow between populations is limited, resulting in evolutionary histories similar to those of macroorganisms. However, population genomic approaches have seldom been applied to this question. Here, we analyse whole-genome resequencing data for all 36 confirmed field isolates of the green alga *Chlamydomonas reinhardtii*. At a continental scale, we report evidence for putative allopatric divergence, between both North American and Japanese isolates, and two highly differentiated lineages within N. America. Conversely, at a local scale within the most densely sampled lineage, we find little evidence for either spatial or temporal structure. Taken together with evidence for ongoing admixture between the two N. American lineages, this lack of structure supports a role for substantial dispersal in *C. reinhardtii* and implies that between-lineage differentiation may be maintained by reproductive isolation and/or local adaptation. Our results therefore support a role for allopatric divergence in microbial eukaryotes, while also indicating that species may be ubiquitous at local scales. Despite the high genetic diversity observed within the most well-sampled lineage, we find that pairs of isolates share on average ~9% of their genomes in long haplotypes, even when isolates were sampled decades apart and from different locations. This proportion is several orders of magnitude higher than the Wright–Fisher expectation, raising many further questions concerning the evolutionary genetics of *C. reinhardtii* and microbial eukaryotes generally.

2.3 Introduction

'Everything is everywhere: but the environment selects' (Baas Becking 1934) has been a long-standing tenet of microbiology (O'Malley 2008). Under this paradigm, dispersal is considered to be effectively unlimited, and the biogeography and evolutionary histories of microbial species should therefore be determined by ecology, rather than geography. For microbial eukaryotes (i.e. protists and other unicellular/colonial eukaryotes), this has been extended to the *ubiquity model* (Finlay and Fenchel 1999; Finlay 2002; Fenchel and Finlay 2004), which predicts both cosmopolitan distributions and low rates of speciation, due to the extremely large population sizes and high dispersal of species. This view has been countered by the *moderate endemicity model* (Foissner 1999; Foissner 2006; Foissner 2008)*,* which posits that dispersal is limited for many species, and as such the taxonomic diversity, biogeography, and evolution of microbial eukaryotes is generally expected to be more similar to that of macroorganisms. Exploring the validity of these opposing models is thus crucial for determining microbial eukaryotic biodiversity, for understanding the rate and mode of speciation in understudied lineages, and for providing insights into the ecology and evolutionary histories of individual species of interest.

Empirical tests of the two competing models have, however, largely been based on morphology, and their interpretation has been highly dependent on the species concept employed (Caron 2009). DNA sequence-based studies of microbial eukaryotes are therefore of great importance, primarily to broadly delineate species (due to the prevalence of cryptic speciation (Lahr et al. 2014)), but more specifically to characterise the nature of population structure within species. Genetic structure can arise as a result of barriers to gene flow formed by limited dispersal (allopatry or isolation by distance), reduced establishment of migrants ('isolation by adaptation'), or more complex patterns caused by founder events ('isolation by colonisation') (Orsini et al. 2013). Exploring the extent of population structure and its causes can be used to test between the *ubiquity* and *moderate endemicity* models, since the former predicts a lack of divergence in allopatry or isolation by distance, and little evidence for recent speciation events, in contrast to what is observed in many plants and animals. Evidence for genetically structured populations has recently been reported across a variety of taxa and habitats, including examples from ciliates (Zufall et al. 2013), amoebae (Douglas et al. 2011; Heger et al. 2013), diatoms (Casteleyn et al. 2010; Sjöqvist et al. 2015; Vanormelingen et al. 2015; Whittaker and Rynearson 2017), dinoflagellates (Lowe et al. 2012; Rengefors et al. 2012), raphidophytes (Lebret et al. 2015), and fungi (Carriconde et al. 2008; Ellison et al. 2011). While many of these studies showed clear evidence for geographical structure (supporting the *moderate endemicity model*), the majority were limited in resolution due to the small number of marker loci used. Microbial eukaryotes remain severely understudied relative to their abundance and phylogenetic diversity (Pawlowski et al. 2012), and currently very few population genomics datasets exist for free-living species (Johri et al. 2017). Such datasets are required to fully capture patterns of genetic diversity within and between populations, to reveal complex patterns of migration and gene flow, and to identify loci putatively contributing to local adaptation and speciation.

As introduced in Chapter 1, despite its importance as a model system the sampling of *C. reinhardtii* is limited and little is known of the species' ecology. For many years, *C. reinhardtii* had only been isolated from eastern North America, suggesting that the species may be endemic (Pröschold et al. 2005). However, isolates that are interfertile with N. American laboratory strains have since been discovered in Japan, implying a more cosmopolitan distribution (Nakada et al. 2010; Nakada et al. 2014). Two previous studies have reported evidence for population structure in field isolates of *C. reinhardtii* (Jang and Ehrenreich 2012; Flowers et al. 2015), but sampling was limited to N. America, and between the studies a total of only 12 isolates were analysed, limiting the inferences that could be drawn. Furthermore, the low number of sequenced isolates has hindered the study of the population genetics of the species.

In this chapter, we analyse whole-genome re-sequencing data for all 36 known *C. reinhardtii* field isolates. We explore patterns of population structure at three scales, (i) local, both between and within sites and time points in Quebec, (ii) within continent, between N. American isolates, and (iii) between continent, specifically between N. American and Japanese isolates. Overall, we report evidence for allopatric divergence, both between N. American and Japanese isolates, and putatively between two highly differentiated lineages in N. America, supporting the *moderate endemicity model* for the species. We find evidence for substantial admixture between the N. American lineages, providing some of the first insights into the ecology and dispersal capability of *C. reinhardtii*. Furthermore, within Quebec we find little signature of strong geographic or temporal structure. Finally, we report the extensive sharing of unexpectedly long genomic tracts likely to have been inherited identical by descent between pairs of isolates at local scales, and discuss several potential causes of this surprising result.

2.4 Results

2.4.1 Whole-genome re-sequencing of *Chlamydomonas reinhardtii* field isolates

Using existing and newly sequenced whole-genome re-sequencing data, we assembled a species-wide sample consisting of 42 isolates, sampled from 11 sites/time points (Figure 1, Table S1 for detailed sampling and sequencing information). Three isolate pairs and one isolate trio, all of which were sampled in Quebec, were found to be clonal (2.7.3, Table S2). Although each isolate was derived from an independent soil sample, all identified clone mates were sampled at the same site and time, which has been observed previously in the case of the clonal pair CC-1952 and CC-2290 (Jang and Ehrenreich 2012). Additionally, CC-3078 was found to be identical to the laboratory strain CC-1010, which was used in mating trials following sampling (Sack et al. 1994) and therefore likely replaced the original isolate at that time. An additional 12 isolates, sampled in Quebec 1993/94, were found not to be *C. reinhardtii* (2.7.4, Table S3). After retaining only one isolate for each clonal pair/trio, the final species-wide dataset comprised 36 isolates and 5.88 million single nucleotide polymorphisms (SNPs), with $\pi_{\text{genome-wide}} = 0.0210$, $\pi_{4D} = 0.0288$, and $\pi_{0D} = 0.00657$ (4D and 0D refer to four-fold and zero-fold degenerate sites, respectively). To our knowledge, this dataset encompasses all genetically unique field isolates of *C. reinhardtii*.

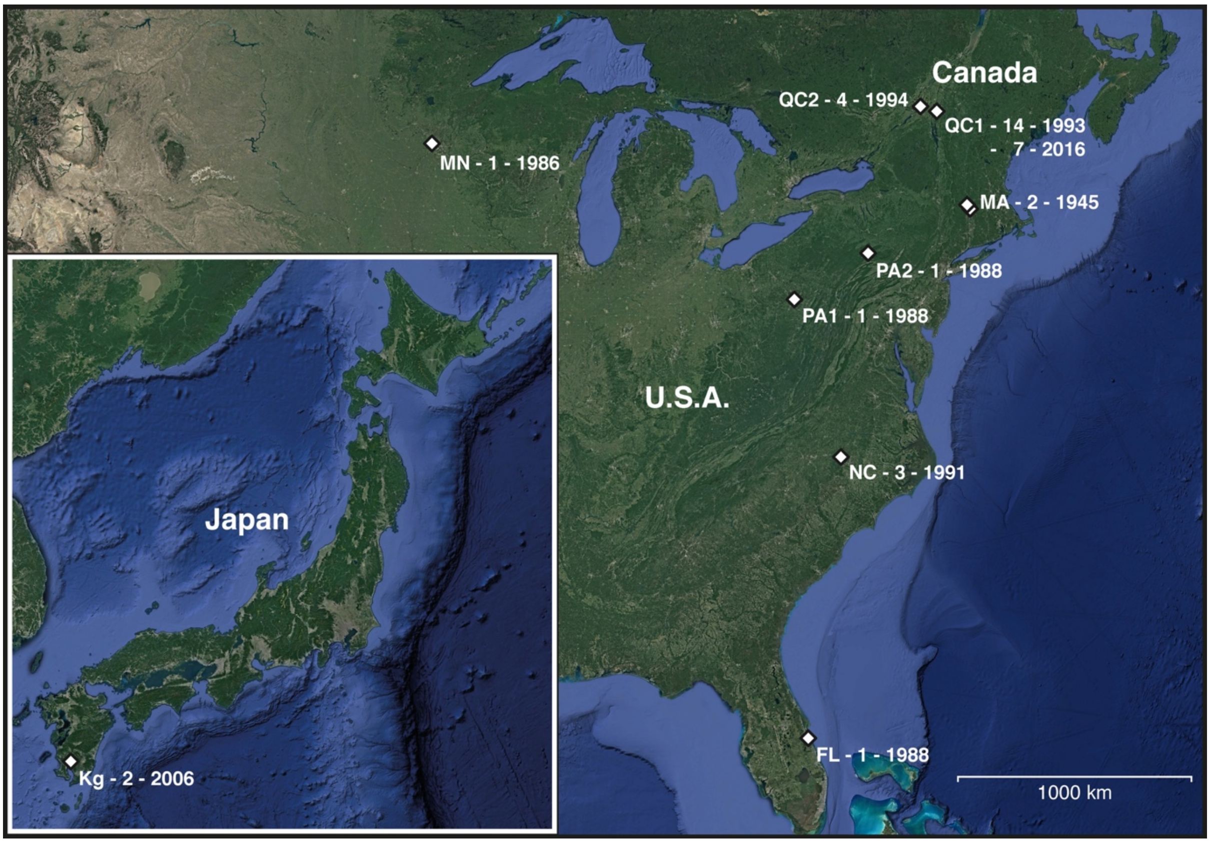


Figure 1. Sampling locations and years for all field isolates included in analyses. Format is 'site – number of isolates – year', where the number of isolates refers to genetically unique (i.e. non-clonal) samples. Location abbreviations are as follows: QC – Quebec, MA – Massachusetts, PA – Pennsylvania, NC – North Carolina, MN – Minnesota, FL – Florida, Kg – Kagoshima Prefecture. Quebec refers to two separate sites, Farnham (QC1, 21 total isolates) and MacDonald College (QC2, four isolates). The Massachusetts isolates are also from two sites ~13 km apart, and one site/isolation is represented by two laboratory strains in the species-wide dataset (2.7.1).

2.4.2 Patterns of continental population structure

Species-wide analyses of population structure indicated that genetic variation in *C. reinhardtii* is geographically partitioned both between N. America and Japan, and within N. America. Both a neighbour joining tree (Figure 2A) and principle component analysis (PCA, Figure 2B) were consistent with all isolates clustering as three distinct lineages, (i) a north eastern N. American lineage (NA1, 27 isolates) comprising the Massachusetts isolates and all Quebec isolates except CC-3079, (ii) an approximately Midwest/Mid-Atlantic/South USA lineage (NA2, eight isolates) comprising all isolates from Pennsylvania, North Carolina, Minnesota and Florida, as well as CC-3079, and (iii) a Japanese lineage (JPN) comprising both isolates from Kagoshima Prefecture, Japan. The N. American lineages were broadly consistent with the two groups described by Jang and Ehrenreich (2012), and our designation of these as NA1 and NA2 follows their previous labelling as group 1 and 2. The geographic distinction between NA1 and NA2 was most clearly shown by the genetic similarity of the Massachusetts and Quebec isolates (sampled ~320-350 km north), relative to the larger genetic distances observed between the Massachusetts isolates and CC-2344 (isolated only ~380 km south west, site PA2 in Figure 1). The grouping of a single Quebec isolate, CC-3079, with NA2, was the only anomaly between these geographic groups, potentially indicating a recent migration event (2.4.4, 2.5.2).

Since *C. reinhardtii* is haploid, to further explore population structure we used the haplotype-based fineSTRUCTURE (Lawson et al. 2012). This approach utilises all variant sites, first using the Chromopainter algorithm to "paint" the chromosomes of every individual (the recipients) as a combination of haplotypes from all other individuals (the donors), so that the sites within each recipient haplotype coalesce most recently with the donor. This information can be plotted as a highly informative coancestry matrix, which summarises the number of haplotypes shared between all donor-recipient pairs. The coancestry matrix produced for the species-wide sample corroborated the neighbour-joining and PCA results, with all isolates sharing many more haplotypes in within-lineage recipient-donor pairs, than in between-lineage pairs (Figure 2C). However, the patterns of haplotype sharing in both between- and within-lineage comparisons were not homogenous. There was evident sub-structure within NA2, with the North Carolina isolates clearly more closely related to each other than to the remaining NA2 isolates. Similar patterns of close relatedness were also evident within NA1 for several Quebec pairs. The between-lineage heterogeneity was indicative of admixture between NA1 and NA2 isolates. Specifically, a subset of NA1 isolates, marked by the dashed blue square in Figure 2C, were the recipients of a greater number of NA2 haplotypes than the remaining NA1 isolates. The NA2 isolates CC-2344 and CC-3079 were the most frequent donors to NA1 isolates, which is notable given that they were sampled in the closest geographic proximity to Massachusetts/Quebec. Additionally, admixture potentially explained the variation on the first principal component of the PCA (Figure 2B), where NA1 axis coordinates were strongly correlated with the estimated proportion of introgressed genome from NA2 (see 2.4.4) ($R = 0.920$, $p < 0.01$).

We also performed a standard STRUCTURE analysis (Pritchard et al. 2000; Falush et al. 2003), the results of which were largely congruent with those of fineSTRUCTURE, the neighbour joining tree and PCA. The optimal K was 2 under the ΔK method (Evanno et al. 2005), although it should be noted that this approach identifies the uppermost level of hierarchical population structure, and that groups containing fewer samples (as for JPN here) are often fitted as mixes of better sampled groups (Lawson et al. 2018). For K=2, NA1 and NA2 clustered as discrete populations with respect to each other, with JPN appearing as admixed (although predominantly NA2-like, Figure S1). For K=3, the three lineages clustered independently, with the majority of NA1 isolates (and particularly the "admixed" subset outlined above) and CC-2344/CC-3079 appearing as admixed between the ancestral

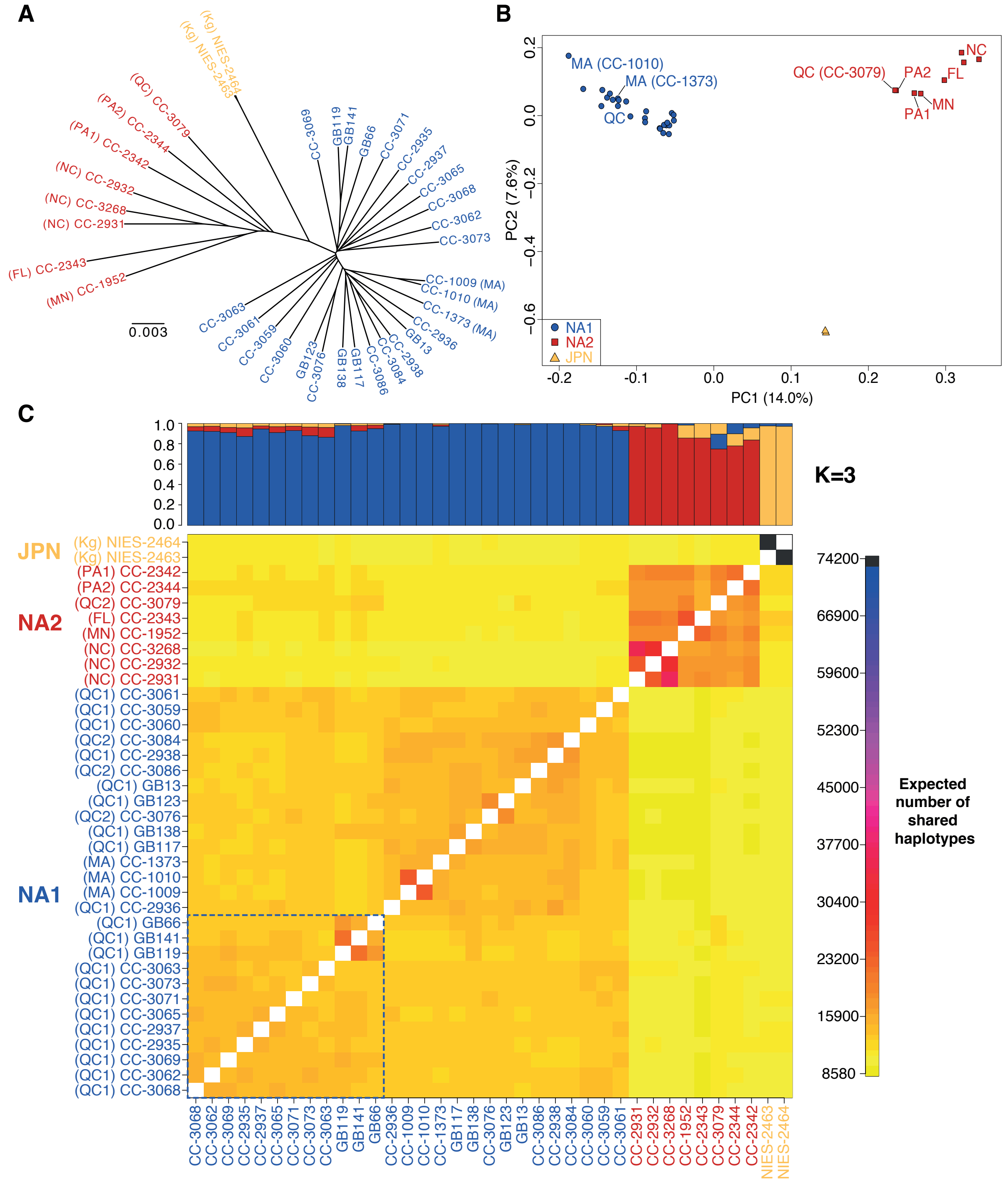


Figure 2. Species-wide population structure analyses.

(A) Neighbour joining tree of all 4D sites, with NA1 isolates coloured blue, NA2 isolates red, and JPN isolates yellow. All nodes had >70% bootstrap support, except for the node connecting CC-3069 with GB119/GB141/GB66.

(B) The first and second axes of the PCA.

(C) fineSTRUCTURE coancestry matrix, in which the colour of the cells represents the expected number of shared haplotypes between donor (columns) and recipient (rows) isolate pairs. The blue dashed square marks a subset of highly admixed NA1 isolates. Sampling locations for each isolate are provided on the y-axis (see Figure 1 for abbreviations). A STRUCTURE plot for three populations is shown above the matrix (see Figure S1 for additional population numbers).

populations corresponding to NA1 and NA2 (Figure 2C, Figure S1). For K=4, this pattern was no longer observed, with no signature of admixture between any of the N. American isolates and JPN, and the genetic variation of NA2 subdivided between two ancestral populations (Figure S1). The sub-structure observed for NA2 at K=4 appeared to divide the North Carolina isolates from the other NA2 isolates, which is not surprising given that there are three isolates from North Carolina, and only one isolate for each of the other sites. For all shown values of K, the majority of NA1 isolates appeared to be partially admixed with NA2.

Finally, there was evidence for isolation by distance between NA2 isolates (Mantel's r^2 = 0.52, p = 0.01), but no significant pattern between NA1 isolates (Figure 3). A pattern of isolation by distance is consistent with the larger geographic range of the NA2 lineage, and the population sub-structure indicated by the fineSTRUCTURE analysis. Given the sparsity of sampling for this group, little can currently be concluded about the extent to which these isolates can be treated as a single evolutionary population.

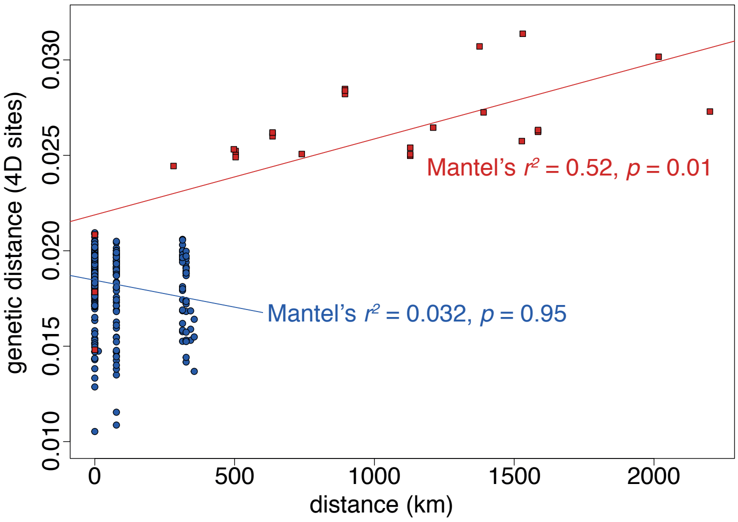


Figure 3. Isolation by distance in NA1 and NA2.
Mantel tests performed on matrices of genetic distance and geographical distance for pairwise comparisons within NA1 (blue) and NA2 (red).

2.4.3 Population structure inferences from the organelle genomes

To explore patterns of population structure using the *C. reinhardtii* organelle genomes, we produced haplotype networks of the mitochondrial genome and plastid coding sequences (CDS). Relative to the nuclear genome (2.4.2), the patterns observed from the mitochondrion (Figure S2A) and plastid (Figure S2B) were less clear. For both organelles, the two Japanese isolates were clearly distinct from the N. American isolates, and isolates within the N. American lineages did exhibit a tendency to group together (e.g. the North Carolina isolates and the Pennsylvania isolate CC-2342). There was also some support for admixture, for example the highly admixed NA2 isolates CC-2344 and CC-3079 had two distinct mitochondrial haplotypes that were both identical to multiple NA1 isolates. Similarly, the highly admixed NA1 isolates GB119 and GB141 grouped closely with the North Carolina isolates and CC-2342 on the plastid network. However, it was not possible to identify ancestral NA1 and NA2 haplotype groups for either organelle, as within each lineage multiple groups of haplotypes were observed. The lack of distinct haplotypes observed for the plastid genome is unsurprising given that it is known to recombine (Dürrenberger et al. 1996; Hasan et al. 2019), and hence it is likely that the plastid genomes of many isolates consist of multiple haplotypes (which may be highly diverged in the case of admixed isolates). Unfortunately, given the short length (~204 kb) and low genetic diversity of the plastid genome (Ness et al. 2016), there was insufficient power to perform similar population structure analyses to those performed on the nuclear genome. Conversely, the origin of such distinct mitochondrial haplotypes is more difficult to explain, given that there is no evidence of recombination in the mitochondrial genome of *C. reinhardtii* (Hasan et al. 2019). Arbitrarily grouping haplotypes together separated by less than ten mutations gives rise to six distinct haplotype groups, two of which are only present in one (CC-3061) or two (CC-1952 and CC-2937) isolates. Additional sampling from new geographic locations (both within N. America and on other continents) will likely be required to further elucidate the origins of these distinct haplotypes, which may potentially indicate the presence of additional unsampled evolutionary lineages of *C. reinhardtii*.

2.4.4 Admixture profiling and the identification of putatively introgressed genomic regions

To further explore the possibility of ongoing admixture between NA1 and NA2, we applied an *ad hoc* approach to identify and visualise putatively introgressed genomic regions derived from admixture between NA1 and NA2 individuals. We identified marker SNPs for NA1 and NA2 as any SNP where the consensus allele differed between the two lineages (2.7.8). The proportions of marker SNPs matching either the NA1 or NA2 consensus alleles for each isolate in 20 kb windows were then plotted as a heat map along each chromosome (chromosome 3 Figure 4A, all chromosomes Figure S3). For all NA1 isolates, large haplotype blocks indicative of recent introgression from NA2 were observed (defined as consecutive windows where the majority of marker SNPs matched the NA2 allele, 2.7.8), and the total proportion of introgressed genome per NA1 isolate ranged from 5.4% to 21.9% (mean 12.7%, Figure 4B). The NA1 isolates designated as highly admixed from the fineSTRUCTURE analysis were found to have significantly more introgressed sequence than the remaining NA1 isolates (means 17.3% and 9.0%, respectively; Wilcoxon rank sum test, $W = 180$, $p = <0.01$), and in practice this categorical division separated the isolates into two groups with less than or greater than 15 Mb of introgressed sequence (~14% of the genome). The mean proportion of introgressed genome for NA2 isolates was lower at 7.7%, with only CC-3079 (17.6%) and CC-2344 (14.9%) exhibiting similarly substantial signatures of admixture. However, this does not necessarily imply that introgression from NA2 to NA1 is more prevalent than in the opposite direction, given that the current sampling of NA2 isolates is so limited, and that highly admixed NA2 populations in close proximity to Massachusetts/Quebec may exist.

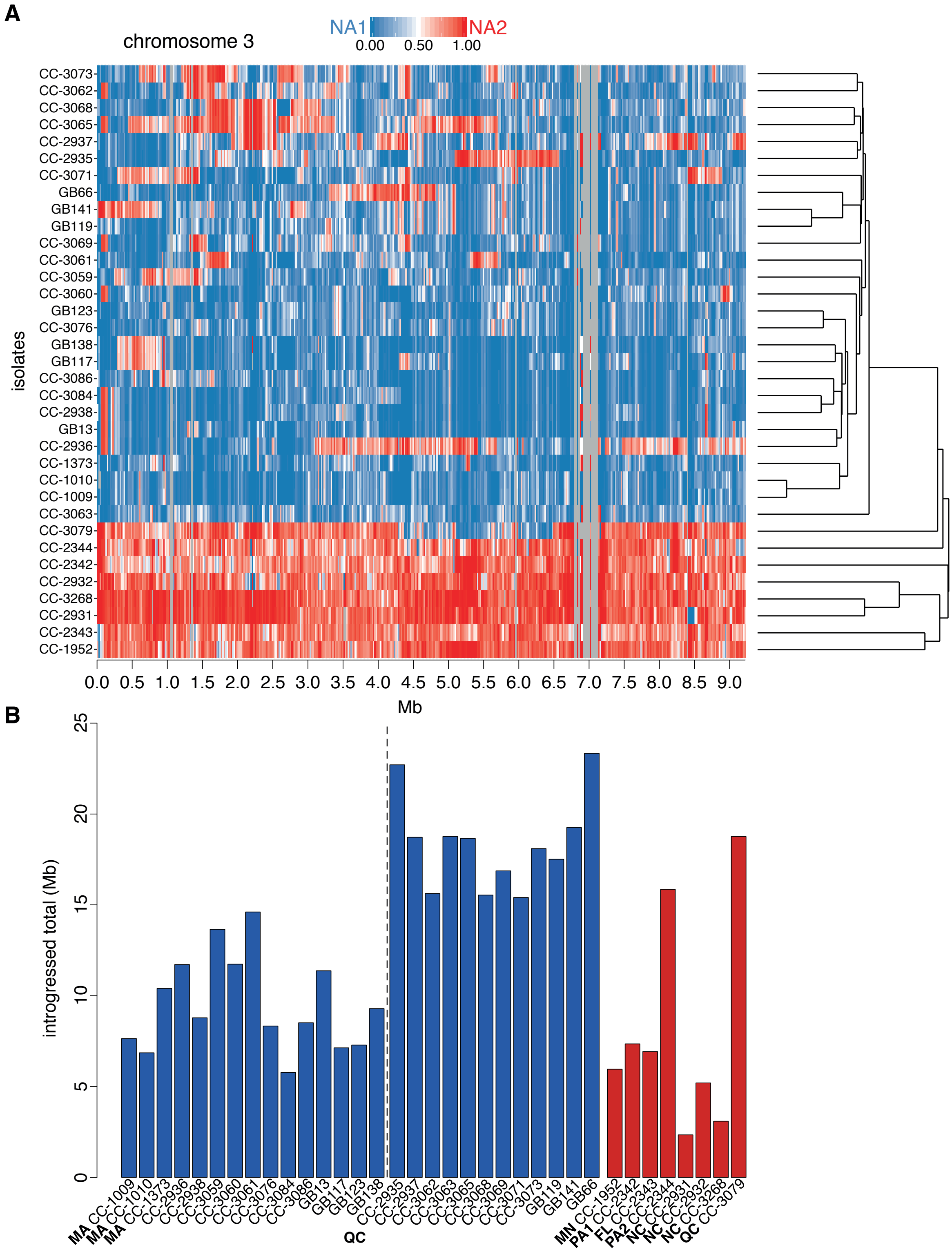


Figure 4. Local ancestry profiling and putative genome-wide introgression.
(A) For each isolate, the proportion of NA1 and NA2 marker SNPs in 20 kb windows is plotted as a heat map along chromosome 3, with 0 (dark blue) representing 100% NA1 SNPs and 1 (dark red) representing 100% NA2 SNPs. Windows containing no sites/SNPs are shown in grey.
(B) Per isolate total of introgressed sequence, with NA1 isolates in blue (bars represent the total length of introgressed sequence from NA2), and NA2 isolates shown in red. The NA1 isolates to the

right of the dashed line are those that were designated as highly admixed from the fineSTRUCTURE analysis (Figure 2C).

A mosaic pattern was observed across the genome of CC-3079 (Figure S4), where on many chromosomes megabase-scale NA1 haplotypes were interspersed on an NA2 genomic background (e.g. chromosomes 3, 4, 6, 7, and 9). However, far shorter transitions between NA1- and NA2-like sequences were also observed, conceivably due to older admixture events. Given that CC-3079 was the only NA2 isolate sampled in Quebec, it is surprising that only 17.6% of the genome was identified as introgressed. Indeed, some chromosomes (e.g. 1, 8, 10 and 16) had no NA1 haplotypes of a size indicative of very recent admixture. Such a pattern of introgression is consistent with at least one admixture event a small number of sexual generations in the past, although assuming all chromosomes undergo at least one crossover per meiosis, the presence of entirely NA2-like chromosomes suggests further mating with NA2 individuals since the putative admixture event(s). From the fineSTRUCTURE analysis, CC-3079 was most closely related to the Minnesota and Pennsylvania isolates, potentially indicating a northern source population from which a migration event could have occurred.

2.4.5 Identity by descent sharing and patterns of local population structure

To quantify relatedness and explore patterns of population structure at a local scale, we used hmmIBD (Schaffner et al. 2018) to identify identical by descent tracts shared between pairs of isolates. The proportion of the genome shared identical by descent between each isolate pair (i.e. the total sharing) was then estimated using three metrics (i) $\hat{\pi}_{IBD}$, the total sharing estimated directly by hmmIBD from the average per-SNP probability of identity by descent (2.7.9), (ii) total sharing for tracts >100 kb, and (iii) total sharing for tracts >500 kb. The estimates differed substantially between metrics, since the absence of shorter tracts in the >100 kb and >500 kb datasets resulted in lower total sharing relative to $\hat{\pi}_{IBD}$ (Table 1, Figure 5A for NA1 only). However, all three metrics were significantly and highly correlated (R = 0.848 – 0.968), and the interpretation of results was consistent across metrics. The following results are given for tracts >100 kb.

As indicated by the fineSTRUCTURE analysis (Figure 2C), there was substantial variation in relatedness between pairs within both NA1 and NA2. Across all NA1 pairs, the distribution of total sharing for tracts >100 kb was approximately normal, although a long tail of the distribution indicated the presence of pairs with a higher genomic fraction of shared tracts (Figure 5A). Total sharing was greater than zero for all 325 NA1 pairs (range 0.3% – 52.0%), and was 9.1% on average, an unexpectedly high figure given the very large effective population size of *C. reinhardtii* (2.4.6, 2.5.6). The variation between isolate pairs may partly be explained by variation in admixture, since introgression is expected to reduce total sharing (Carmi et al. 2013). As expected under this scenario, the cohort-averaged sharing (a per isolate identity by descent summary statistic, 2.7.9) for NA1 isolates was significantly negatively correlated with the inferred proportion of introgressed genome from NA2 (R = -0.675, $p < 0.01$). There was no signature that identical by descent tracts were highly concentrated in particular genomic regions, as ~99% of the genome was included in at least one pairwise tract, and the distribution of the average sharing across all NA1 pairs in 100 kb chromosomal windows was approximately normal (Figure 5B).

Table 1. Average genomic proportions shared identical by descent for isolate pairs within and between samples.

Population/Comparison	$\hat{\pi}_{IBD}$	**Average total sharing >100 kb tracts (%)**	**Average total sharing >500 kb tracts (%)**	π_{4D}	**Number of isolate pairs**
NA1	23.6	9.11	2.64	0.0236	325
Massachusetts	36.2	16.9	3.50	0.0188	1
Quebec	23.4	9.18	2.78	0.0237	276
Farnham 1993	22.2	7.13	1.18	0.0242	91
MacDonald College 1994	35.2	20.8	10.2	0.0193	3
Farnham 2016	29.3	17.3	9.04	0.0218	21
Massachusetts – Quebec	24.3	8.55	1.76	/	48
Farnham 1993 - MacDonald College 1994	23.2	8.31	2.52	/	42
Farnham 1993 - Farnham 2016	21.8	7.99	2.00	/	98
NA2	9.41	2.77	0.959	0.0306	28
North Carolina	32.6	23.2	8.95	0.0190	3
NA2 between locations	0.0595	0.217	0.00	/	12

Proportions of the genome shared identical by descent (i.e. total sharing) are shown for the total predicted by hmmIBD ($\hat{\pi}_{IBD}$), for tracts >100 kb, and for tracts > 500 kb. The number of isolate pairs refers to the total number of pairwise comparisons contributing to the average total sharing. For each lineage, average total sharing is shown for the subsets of isolates discussed in 2.4.5 (e.g. North Carolina for NA2), and comparisons between subsets are labelled as the two subsets separated by a hyphen (e.g. Farnham 1993 – Farnham 2016). For the NA2 between locations comparison all pairwise comparisons within North Carolina were excluded.

Given the prevalence of identity by descent tracts in NA1, it is unclear to what extent total sharing can be used as a proxy for relatedness. Nonetheless, following the assumption that the total sharing is at least partially indicative of the relatedness between a pair of isolates, this relationship can be used to explore local population structure within NA1, and specifically within Quebec. If genetic diversity is spatially or temporally structured at local scales in *C. reinhardtii*, it is expected that total sharing would be higher for within-site isolate pairs (Farnham and MacDonald College, ~80 km apart) relative to between-site pairs, and for within-time point pairs at the same site (Farnham 1993 and 2016) relative to between-time point pairs. There was, however, no support for either of these relationships, with no difference in total sharing for within-site pairs relative to between-site pairs (Wilcoxon rank sum test, $W = 2228$, $p = 0.23$), and no difference for within-time point pairs relative to between-time point pairs (Wilcoxon rank sum test, $W = 5859$, $p = 0.40$). Moreover, there was also no difference in total sharing for pairs within Quebec and Massachusetts, relative to pairs between Quebec and Massachusetts (Wilcoxon test rank sum test, $W = 7054$, $p = 0.50$), where the isolates were sampled ~320-350 km and ~50-70 years apart. Therefore, taken together with the lack of isolation by distance (Figure 3), there appears to be no strong signal of population structure within the current sampling of NA1.

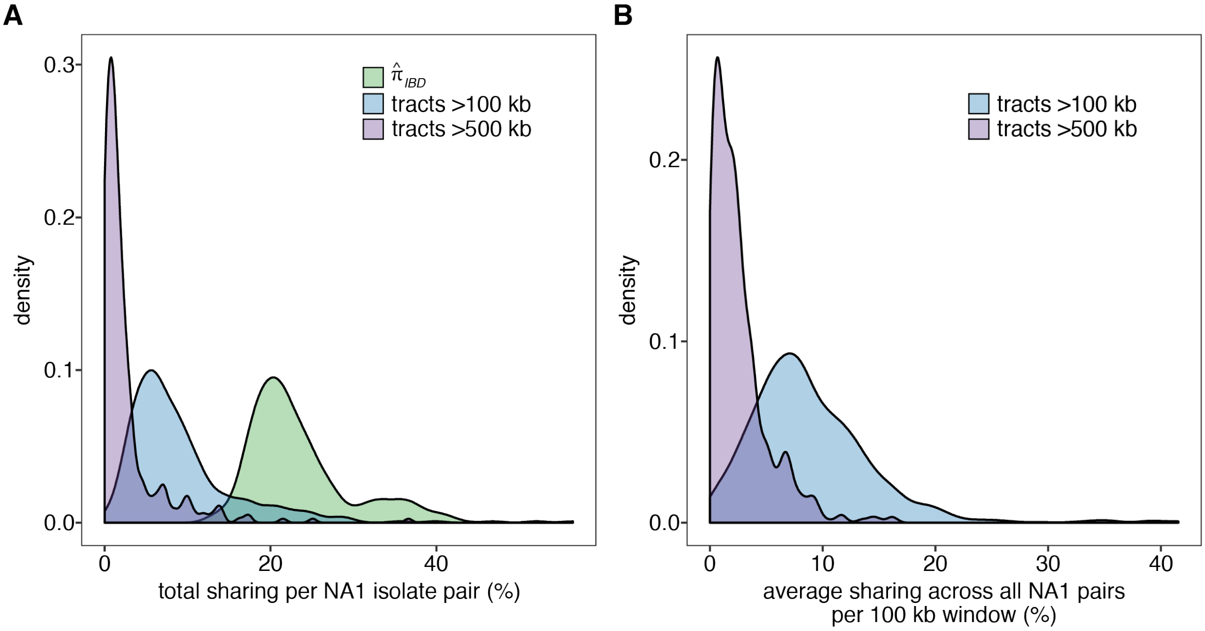


Figure 5. Identity by descent sharing in NA1.
(A) Density plot showing the distributions of estimates of total sharing across all 325 isolate pair comparisons for NA1, shown for the three definitions of identity by descent.
(B) Density plot showing the distribution of the mean sharing across all 325 NA1 pairs per 100 kb genomic window, shown for tracts >100 kb and >500 kb.

Conversely, there were differences between the samples, with the average total sharing within MacDonald College 1994 (20.8%) and Farnham 2016 (17.3%) more than twice that of Farnham 1993 (7.1%). Samples with greater average total sharing exhibited lower putatively neutral genetic diversity (π_{4D}), resulting in the unexpected observation that diversity was marginally higher within a single sample (Farnham 1993 π_{4D} = 0.0242) than within the entire sampled lineage (NA1 π_{4D} = 0.0236, Table 1). The lower average total sharing within Farnham 1993 may be explained by an increased rate of admixture within this sample, as the average proportion of introgressed genome was higher (14.8%) relative to MacDonald College 1994 (7.1%) and Farnham 2016 (12.8%) (Figure 4B). The Farnham 1993 isolate pairs make up the majority of the within-sample pairs in the above within vs between sample statistical comparisons, so the reduction in total sharing for this sample may explain the reported lack of significance. Regardless of this, the average total sharing between Farnham and MacDonald College (8.3%), and between Farnham 1993 and 2016 (8.0%), remain far greater than would be expected if there was strong spatial or temporal structure within Quebec.

In contrast to NA1, there was very little signature of close relatedness between NA2 isolates from different locations. Total sharing for between location NA2 pairs was only 0.2% on average (Table 1), corroborating the presence of population sub-structure in the lineage.

However, within the North Carolina sample (the only site with more than one NA2 isolate), the average total sharing was 23.2%. Taken together with the results for NA1, the independent finding of very high total sharing between North Carolina isolate pairs suggests that *C. reinhardtii* haploid individuals may generally share a substantial proportion of their genomes identical by descent at local scales.

2.4.6 Genetic diversity within lineage, and genetic differentiation and divergence between lineages

Genetic diversity varied substantially between lineages (Figure 6A), with π_{4D} estimates of 0.0236, 0.0306, and 0.00123 for NA1, NA2, and JPN, respectively. Based on these estimates of putatively neutral diversity and a SNP mutation rate of 9.63 x 10^{-10} per site per generation estimated by re-sequencing of *C. reinhardtii* mutation accumulation lines by Ness et al. (2015), the estimated effective population sizes (N_e) for each lineage were 4.91 x 10^7 (NA1), 6.35 x 10^7 (NA2), and 2.56 x 10^6 (JPN) (following $\pi = 2N_e\mu$.) Thus, at least for the N. American lineages, these estimates are consistent with *C. reinhardtii* genetic diversity being amongst the highest reported in eukaryotes (Leffler et al. 2012). It is difficult to conclude to what extent the higher diversity of NA2 relative to NA1 reflects sampling history, since the NA2 isolates have been sampled over a far larger area with generally only one isolate per site (except for the three North Carolina isolates). Indeed, considering single sampling locations, π_{4D} estimated for only the three North Carolina isolates was 0.0190, lower than that calculated for the Farnham 1993 isolates (0.0242), and marginally lower than that for the three MacDonald College NA1 isolates (0.0193), which have a comparable incidence of identity by descent sharing to the North Carolina isolates (Table 1).

Strikingly, genetic diversity for JPN was an order of magnitude lower than that for the N. American lineages, with the estimated π_{4D} of 0.00123 approximately 19 and 25 times lower than the estimated values for NA1 and NA2, respectively. Although based only on two isolates, this did not appear to be an artefact caused by high relatedness. Firstly, the isolates are of opposite mating types, and so are certainly not clonal. Secondly, genetic diversity appeared to be uniformly lower across the genome relative to N. American isolates, with no obvious long invariant tracts as observed for pairs of NA1 isolates (Figure 6B). Indeed, even for the extreme of highly related isolate pairs (e.g. GB119 and GB141, sharing ~50% of their genomes), and for the laboratory strains CC-1009 and CC-1010 (sharing ~75% of their genomes), pairwise genetic distances greatly exceeded that observed between the two JPN isolates, as shown by the branch lengths of the neighbour joining tree (Figure 2A).

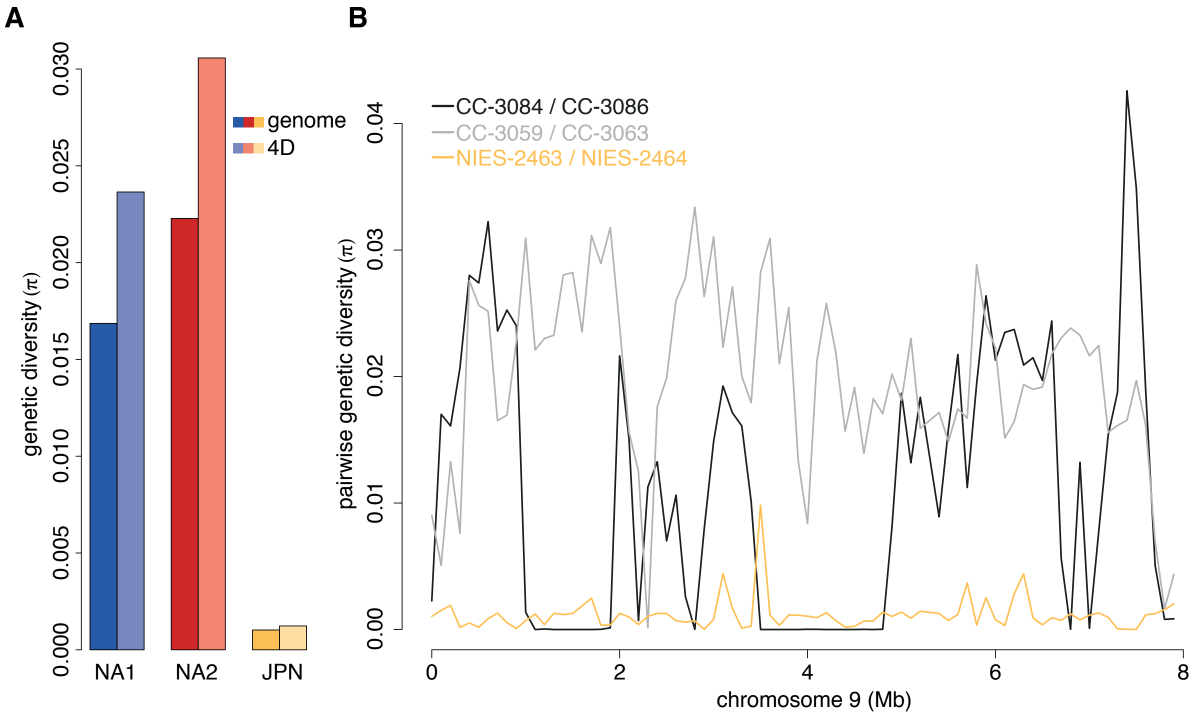


Figure 6. Summary of genetic diversity within *C. reinhardtii* lineages.
(A) Genome-wide and 4D within-lineage genetic diversity for NA1, NA2 and JPN.
(B) A comparison of pairwise genetic diversity estimated along chromosome 9 in 100 kb windows, for the JPN isolates, and for Quebec isolate pairs exhibiting a low (CC-3059 – CC-3063) and high (CC-3084 – CC-3086) incidence of identity by descent sharing.

The NA1 and NA2 lineages were highly differentiated, both genome-wide (F_{st} = 0.25) and at 4D sites (F_{st} = 0.24) (Table 2). Only 30.6% of the 7.19 million SNPs segregating in the N. America sample were shared between the lineages, with 37.3% private to NA1, and 31.8% private to NA2. Results were similar for 4D SNPs, with a slightly higher percentage shared between the lineages (33.0%). Despite the majority of SNPs being private to either lineage, only 0.3% (genome-wide) and 0.2% (4D) of SNPs were fixed, consistent both with admixture and the expected weak force of genetic drift due to the high effective population size of the species. The average number of pairwise differences between the lineages (d_{xy}) was estimated as 0.0274 (genome-wide) and 0.0364 (4D), and thus two sequences drawn randomly between NA1 and NA2 contained 54.2% more differences than two NA1 sequences, and 19.0% more differences than two NA2 sequences (for 4D sites, based on comparison to within-lineage π_{4D}). After masking introgressed regions for both lineages, the overall percentage of shared SNPs decreased to 19.8% and 22.6%, F_{st} increased to 0.34 and 0.32, and d_{xy} increased to 0.0281 and 0.0374 (all for genome-wide and 4D sites, respectively). Surprisingly, the JPN lineage was no more genetically distant from NA1 (4D d_{xy} = 0.0343) and NA2 (4D d_{xy} = 0.0376), than NA1 and NA2 were from each other.

Table 2. Differentiation and divergence between *C. reinhardtii* lineages (NA1 26 isolates, NA2 eight isolates, JPN two isolates).

		NA1 - NA2	**NA1- JPN**	**NA2 - JPN**	**NA1 - NA2 (introgression masked)**
SNPs	genome-wide	7,188,929	4,496,586	4,167,903	6,379,381
	4D	881,984	598,261	562,782	798,407
shared (%)	genome-wide	30.6	0.279	0.222	19.8
	4D	33.0	0.315	0.261	22.6
private A (%)	genome-wide	37.3	88.9	84.4	36.1
	4D	36.5	90.1	85.7	35.6
private B (%)	genome-wide	31.8	1.22	1.32	42.0
	4D	30.4	1.00	1.12	40.1
fixed (%)	genome-wide	0.301	9.67	14.0	2.21
	4D	0.194	8.60	12.9	1.64
F_{st}	genome-wide	0.25	0.64	0.59	0.34
	4D	0.24	0.63	0.58	0.32
d_{xy}	genome-wide	0.0274	0.0256	0.0283	0.0281
	4D	0.0364	0.0343	0.0376	0.0374

For private SNPs, A is the first lineage in the comparison, and B the second. Introgression masked refers to the NA1 – NA2 comparison after removing genomic regions identified as introgressed for each individual.

2.5 Discussion

2.5.1 Overview

In this study we have used genome-wide data to explore patterns of population structure across field isolates of *C. reinhardtii.* Taking advantage of the haploid state of the isolates, we applied haplotype-based analyses to characterise structure at both continental and local scales, and to infer patterns of admixture between the two identified N. American lineages. In what follows, we contextualise these findings within the ongoing debate concerning the nature of biogeography and speciation in microbial eukaryotes, and discuss further insights concerning the evolutionary history and ecology of *C. reinhardtii*. Finally, we discuss the surprising prevalence of identity by descent sharing between isolates sampled at local scales.

2.5.2 The North American biogeography of *Chlamydomonas reinhardtii*

Based on current sampling, the evidence for three geographically distinct lineages of *C. reinhardtii* contradicts the predictions of the *ubiquity model*, under which little geographic population structure is expected. Interestingly, there are notable similarities between the observed biogeography of *C. reinhardtii* and the best studied microbial eukaryote in this context, *Saccharomyces paradoxus.* This wild yeast has been shown to form a species complex, comprising highly differentiated lineages on different continents, suggesting allopatric divergence and speciation (Koufopanou et al. 2006; Kuehne et al. 2007; Liti et al. 2009). Within N. America, two allopatric lineages of *S. paradoxus* have been described, which exhibit signatures of local adaptation and reproductive isolation characteristic of incipient species (Charron et al. 2014; Leducq et al. 2014; Leducq et al. 2016). Similar to *C. reinhardtii*, one lineage has a more restricted range in the north east, while the other is widely distributed to the south and west, with a sympatric zone occurring along Lake Ontario and the St. Lawrence River (Charron et al. 2014). This biogeography is consistent with allopatric divergence in the Atlantic and Mississippian glacial refugia during the last glacial maximum (~110,000 – 12,000 year ago), which has been documented in numerous plants and animals (Charron et al. 2014). Thus, although as a morphological entity *S. paradoxus* fulfils the '*everything is everywhere'* maxim, it in fact consists of several cryptic species that have undergone allopatric speciation events, including a putative event in glacial refugia contemporaneous with several plants and animals.

Whether glacial refugia can explain the biogeography of the two N. American *C. reinhardtii* lineages will largely be contingent on further sampling, especially in what would be expected to be the north eastern limits of the NA2 range (i.e. south west of New England and the St. Lawrence River). However, the observed biogeography is consistent with such a scenario, under which NA1 would have persisted in the Atlantic refugium (located east of the Appalachians), before re-colonising Massachusetts and Quebec. This could also explain the sub-structure observed for NA2, which may have a markedly different evolutionary history to NA1, with the possibility of multiple refugia (e.g. Mississippian, Virginia/Carolinas Atlantic coast, and further south) connected by varying amounts of gene flow at different times. Furthermore, the two lineages cannot easily be explained by climate or other environmental factors, since NA2 includes both one of the most northerly (CC-1952, Minnesota) and the most southerly (CC-2343, Florida) isolates, and the Massachusetts and Pennsylvania sites presumably share similar environments. However, we have not explicitly tested any environmental variables in this study, and this will form an important aspect of future research.

That essentially all NA1 isolates exhibit signatures of admixture with NA2 individuals supports a role for substantial dispersal in *C. reinhardtii*. Given that the length of the observed introgressed haplotypes are considerably longer than the physical distance over which linkage disequilibrium (LD) decays in the species (~10-20 kb (Flowers et al. 2015; Hasan and Ness 2020)), admixture is likely to have occurred in the relatively recent past. Furthermore, that a single highly admixed NA2 isolate (CC-3079) was present within our small Quebec sample suggests that both migration and gene flow are ongoing. Under such a scenario, the findings that the two lineages remain so highly differentiated in the face of migration and gene flow potentially indicates the presence of reproductive isolation and/or local adaptation. Although all three identified lineages cross successfully in the laboratory (Pröschold et al. 2005; Nakada et al. 2014), the crossing success of the Florida and Pennsylvania isolates with laboratory strains has been reported to be reduced relative to that between laboratory strains (Spanier et al. 1992). There are also substantial phenotypic differences between isolates (Flowers et al. 2015), and it should be possible to re-visit such variation in the context of the two N. American lineages, and to further test for reproductive isolation in the laboratory (e.g. via fitness assays of ‘hybrid’ progeny).

The mosaic genome of CC-3079 also provides further insights into the ecology of *C. reinhardtii*. The observed pattern cannot simply be explained by an NA2 migrant arriving in Quebec and subsequently mating with only NA1 individuals, since several chromosomes show no signature of recent introgression, implying that mating between other NA2 individuals occurred after the inferred admixture event(s). This could be explained if CC-3079 were itself a migrant from an unsampled location in which both NA1 and NA2 individuals occur in sympatry and hybridise. Alternatively, an NA2 ancestor of CC-3079 may have migrated to Quebec, implying the presence of other NA2 individuals at the site. Almost nothing is known about the dispersal capability and mechanisms in *C. reinhardtii*, although there is abundant evidence for the passive dispersal of dormant propagules (such as the *C. reinhardtii* zygospore) of various species (De Meester et al. 2002). As such propagules are resistant to environmental stresses, they can be transported over long distances via biotic (e.g. birds and insects), abiotic (e.g. wind and water), or anthropomorphic vectors. Additionally, as *C. reinhardtii* zygospores adhere to each other (Harris 2009), a single migration event may have the potential to introduce many migrant individuals of both mating types, which could explain the implied presence of other NA2 individuals at the sampling site.

2.5.3 The Japanese isolates and the wider biogeography of *Chlamydomonas reinhardtii*

Although the evolutionary history of the Japanese isolates is essentially unresolved based on current sampling, their inclusion in this study at least indicates that *C. reinhardtii* on different continents may be expected to form substantially divergent lineages. However, under a model of allopatric divergence between N. American and Japanese *C. reinhardtii,* it is surprising that the JPN lineage is no more genetically distinct from either NA1 or NA2, than NA1 and NA2 are from each other. One speculative explanation is that the Japanese isolates were derived from a third unsampled N. American lineage that underwent divergence from NA1 and NA2 simultaneously (e.g. in Pacific or Beringian refugia), before migration to Japan. Water birds are thought to be a major mechanism of algal dispersal (Kristiansen 1996), and western N. America, and in particular Alaska, is linked to Japan by the flyways of several migratory bird species. Alternatively, gradual dispersal across the Bering land bridge could also give rise to a similar pattern, leading to the prediction that any East Asian and Alaskan *C. reinhardtii* may be genetically similar. The strikingly low genetic diversity of the two Japanese isolates relative to the N. American lineages is also surprising. If the lineage was established from a larger population by migration (which could in principle occur from a single zygospore), then such a founder effect would be expected to reduce diversity via a severe bottleneck (De Meester et al. 2002). Supporting this hypothesis, any population present in Kagoshima must be geologically young, as a result of the formation of the Aira Caldera ~30,000 years ago, and the Akahoya eruption ~7,000 years ago (Machida and Arai 2003).

As a result of the historic difficulty in isolating *C. reinhardtii* (Pröschold et al. 2005), it is likely that the current sampling primarily reflects the distribution of researchers. Intercontinental distributions of more conspicuous Volvocalean algae have been documented (e.g. Kawasaki et al. (2015)), and given the geographic distance between eastern N. America and Japan, it would not be surprising if *C. reinhardtii* is shown to have a considerably wider distribution in the future. However, far more extensive sampling across multiple regions and habitats, alongside improvements in sampling methodology, will be required to address this.

2.5.4 Patterns of population structure and genetic diversity at a local scale

In facultatively sexual organisms, under certain conditions clonal erosion can generate population structure and reduce genetic diversity at local scales (Vanoverbeke and De Meester 2010). Prior to this study, almost nothing was known about the local structure of genetic diversity in *C. reinhardtii,* and it was unknown whether a single site would be dominated by clonal lineages. Although our sample contained a small number of clonal pairs/trios, most isolates sampled at single sites were genetically distinct, and diversity at single sites and time points was of the same magnitude as the total lineage diversity. Although the extent of identity by descent sharing appeared to vary between sites and time points in Quebec, we found no evidence for strong population structure at this scale. The lack of structure observed in space further supports the considerable dispersal potential of *C. reinhardtii*. The lack of structure observed in time could potentially be explained by long-term zygospore dormancy, which would result in isolates sampled many years apart being separated by far fewer sexual generations than would otherwise be expected. Such a phenomenon is known in other chlorophyte algae, where dormant zygospores are capable of forming propagule banks (Fryxell 1983), and it is known that *C. reinhardtii* zygospores are resistant to both long-term freezing and desiccation (Harris 2009). Propagule banks have also been hypothesised to contribute to high levels of genetic diversity, as populations can be re-seeded with haplotypes present at previous time points (Rengefors et al. 2017; Shoemaker and Lennon 2018), and therefore long-term zygospore dormancy could be a contributing factor to the high diversity estimated for *C. reinhardtii*.

C. reinhardtii population genetics analyses have been hindered by the absence of a suitable set of isolates, and the lack of understanding as to what constitutes a ‘population’ in the species. The high genetic diversity found at single sites in this study now presents the opportunity to use samples from single sites (e.g. Farnham 1993) for future analyses. Furthermore, given the lack of structure between sites/time points, the entire Quebec sample could conceivably be analysed together. Although the extent of identity by descent sharing between these isolates requires further explanation (2.5.6), the delineation of a group of isolates suitable for population genetics analyses has the potential to greatly enhance the use of *C. reinhardtii* in evolutionary biology research.

2.5.5 Broader perspectives on microbial biogeography and speciation

Taken together with the evolutionary history of *S. paradoxus*, our interpretation of *C. reinhardtii* continental population structure supports a role for allopatric differentiation (and potentially speciation) in microbial eukaryotes. This permits the rejection of the *ubiquity model* in these cases, supporting the more similar rates of speciation between microbial eukaryotes and macroorganisms predicted by the *moderate endemicity model*, and implying that microbial species may be far more speciose than existing taxonomic descriptions suggest. It is worth noting, however, that the *moderate endemicity model* does not predict frequent allopatric speciation (instead favouring various forms of non-allopatric speciation) (Foissner 2008), and in this sense the model may need to be revised. De Meester et al. (2002) detailed the role of glacial refugia in speciation events for various zooplankton, and it may be that similar allopatric events are also commonplace in microbial eukaryotes. However, it is unclear to what extent the results for two terrestrial species can be extrapolated, and the exploration of similar patterns across a far larger range of species is obviously required to fully address this question.

2.5.6 Identity by descent sharing between *Chlamydomonas reinhardtii* isolates

The original motivation for identifying identical by descent tracts was to quantify between-pair relatedness and explore patterns of local population structure. However, the most surprising result of these analyses was that on average a pair of NA1 isolates share 9.1% of their genomes in tracts >100 kb, and that an even higher proportion was independently observed between the three isolates sampled in North Carolina. Even more unexpectedly, isolates from Massachusetts and Quebec (sampled ~50-70 years apart) share 8.6% of their genomes identical by descent on average. This highlights a striking dichotomy: how can essentially the entire sampled population appear to share recent ancestry, yet genetic diversity be maintained at a high level? Although much of our understanding of identity by descent in populations has been built upon pedigrees (Thompson 2013), population-level theory has recently been developed for tracts defined by arbitrary genetic length cut-offs (Palamara et al. 2012; Carmi et al. 2013; Carmi et al. 2014). Using equation 4 of Carmi et al. (2013), and based on the estimated N_e for NA1 and a minimum tract length of 100 kb (~1.2 cM), the average proportion of the genome shared identical by descent between a pair of individuals in a Wright-Fisher population is expected to be ~0.00017%, four orders of magnitude lower than observed.

Although we currently lack an explanation for this discrepancy, there are several possibilities that can currently be considered. First, *C. reinhardtii* evidently does not meet the assumptions of a Wright-Fisher population, and therefore a stochastic process may be responsible. Clonal reproduction is expected to result in a high variance in reproductive success (Tellier and Lemaire 2014), and zygospore dormancy would result in overlapping generations, although further theoretical work will be needed to address the effects of such processes on identity by descent. Second, it is conceivable that many long shared genomic tracts could arise in a population as a result of pervasive positive selection combined with long-range effects of selection on linked sites. Frequent adaptive evolution and the resulting effects of hitchhiking on linked sites has recently been evoked to explain the low observed diversity in the ubiquitous phytoplankton species *Emiliania huxleyi* (Filatov 2019). Although *C. reinhardtii* obviously differs from this case with respect to genetic diversity, if pervasive positive selection acted mostly on standing variation in the species, it is possible that soft selective sweeps could result in multiple haplotypes rising to high frequency, while maintaining high genetic diversity. Third, if there is a high diversity of structural variants segregating in *C. reinhardtii* populations there may be recombination suppression between certain haplotypes. Physical recombination has only been studied between a very small number of *C. reinhardtii* isolates (Kathir et al. 2003; Liu et al. 2018), and additional experimental work will be required to further explore recombination in the species. In a broader sense, empirical studies of other species with similar life cycles will also be crucial to determining the generality of this result. Similar patterns may easily have been missed in diploid species due to a lack of phased haplotypes.

2.6 Conclusions

C. reinhardtii is divided into three geographically distinct lineages based on current sampling, supporting the *moderate endemicity model* of microbial eukaryote biogeography. *C. reinhardtii* is likely to have substantial dispersal capability, implying that reproductive isolation and/or local adaptation may be maintaining genetic differentiation between the two N. American lineages in the face of ongoing migration and gene flow. High dispersal may also prevent the evolution of population structure at local geographic scales. Within two independent populations an extremely high incidence of identity by descent sharing was observed, raising several interesting questions regarding the evolutionary genetics of *C. reinhardtii*.

2.7 Methods

2.7.1 Sampling and whole-genome re-sequencing

Sampling and whole-genome re-sequencing of the field isolates available from the Chlamydomonas Resource Centre (https://www.chlamycollection.org) has mostly been described previously. Briefly, sequencing data for 11 isolates sampled at eight locations between 1945 and 1994 were produced by Flowers et al. (2015), with the exception of CC-2932 (Jang and Ehrenreich 2012). We obtained and sequenced the isolate CC-3268, since it was not included in previous studies. A total of 31 isolates (CC-3059 – CC-3089 in the collection), sampled in 1993/94 from two sets of fields ~80 km apart in Quebec (Farnham and MacDonald College), were first screened by Sanger sequencing of introns VI and VII of the *YPT4* gene, which are species-specific markers in volvocine algae (Liss et al. 1997). Eighteen isolates were confirmed as authentic *C. reinhardtii*, sequencing of which was described by Ness et al. (2016). A further eight previously undescribed isolates (referred to as

GB# in this study) were sampled from Farnham in 2016, using the protocol of Sack et al. (1994).

Data produced by Gallaher et al. (2015) for the laboratory strains CC-1009 and CC-1010 were also included. Since all laboratory strains are derived from a single zygospore sampled in Massachusetts in 1945, the genomes of these strains consist of two parental haplotypes, although across all strains ~75% of the genome appears to have originated from one parent (Gallaher et al. 2015). CC-1009 and CC-1010 have inherited opposite parental haplotypes, and so together maximise the genetic variation present amongst the laboratory strains. Both strains were included in the analyses of population structure and admixture, where they can be analysed as genetically distinct at ~25% of genomic sites. For analyses where the independence of isolates was required (i.e. the calculation of population genetics statistics and the identification of identity by descent tracts), CC-1009 was excluded.

For the 2016 Farnham isolates and CC-3268, DNA was extracted by phenol-chloroform extraction following Ness et al. (2012). Whole-genome re-sequencing was performed on the Illumina HiSeq 2000 platform (100 bp paired-end reads) for the Farnham isolates, and on the Illumina Hiseq 4000 platform (150 bp paired-end) for CC-3268, both at BGI Hong Kong. The modified PCR conditions of Aird et al. (2011) were used during library preparation to accommodate the high GC-content of *C. reinhardtii* (mean nuclear GC = 64.1%). The Japanese isolates NIES-2463 and NIES-2464 were sequenced using the Illumina MiSeq platform (300 bp paired-end). Detailed sampling history and sequencing metrics for all isolates are provided in Table S1.

2.7.2 Read mapping and variant calling

Read mapping and initial variant calling were performed as described by Ness et al. (2016). Briefly, reads were mapped to v5 of the *C. reinhardtii* reference genome (Blaby et al. 2014) using BWA-MEM v0.7.5a-r405 (Li 2013) with default settings. The plastid (NCBI accession NC_005353) and mitochondrial (NCBI accession NC_001638) genomes were appended to the reference, as was the MT^{-} locus (NCBI accession GU814015), since the reference genome isolate is MT^{+}. Genotypes were called using the GATK v3.5 (DePristo et al. 2011) tool HaplotypeCaller, and the resulting per isolate Genomic Variant Call Files (gVCF) were combined to a species-wide Variant Call File (VCF) using GenotpyeGVCFs with the following non-default settings: sample_ploidy=1, includeNonVariantSites=true, heterozygosity=0.02, indel_heterozygosity=0.002.

Only invariant and biallelic sites were considered for analyses. Filters were applied independently on the genotype calls of each isolate, as opposed to per site. Retained genotypes required a minimum of three mapped reads, with the total depth not exceeding the average depth for the isolate in question plus four times the square root of the average depth (to remove regions with copy number variation (CNV) (Li (2014)). Genotypes located 5 bp either side of an indel were filtered, to avoid false positives due to misaligned reads. SNPs with a genotype quality (GQ) <20, or with <90% of the informative reads supporting the called genotype, were filtered. All sites from the ~600kb *MT*$^+$ (between the *NIC7* and *THI10* genes (De Hoff et al. 2013)) and *MT*$^-$ loci were filtered. For the population structure analyses no missing genotype data were allowed, resulting in the analysis of 1.44 million SNPs. For analyses comparing the different identified *C. reinhardtii* lineages, to maximise the number of callable sites a minimum of 50% of isolates within each lineage were required to have genotypes that passed filtering (except for the Japanese isolates, where both were required), resulting in the analysis of 58.0% of sites genome-wide (61.77 Mb) and 74.4% of 4D sites (6.18 Mb).

2.7.3 Identification of clonal isolates

Clonal pairs/trios of isolates were identified based on the extremely low number of called variants observed between the isolates in question (Table S2). The SNPs that were called for each pair/trio were manually checked using IGV v2.5.2 (Robinson et al. 2011). Almost all called SNPs appeared to be heterozygous when viewed as read alignments, and so were clearly false positive calls (since *C. reinhardtii* is haploid). This is most likely due to CNV between the reference genome strain and the isolates in question. In support of this, the lowest number of called SNPs in a clonal pair (245) was observed for the laboratory strain CC-1010 and its clone CC-3078, which are closely related to the reference genome and would be expected to harbour fewer CNVs. Conversely, the highest number (1,680) was observed for the clonal pair CC-3075/CC-3079, which are the most genetically distant from the reference genome of any of the clonal pairs/trios (2.4.2) and are therefore the most likely to harbour a higher number of CNVs.

2.7.4 Quebec isolates found not be *Chlamydomonas reinhardtii*

After PCR amplification and Sanger sequencing of the highly variable markers *YPT4* introns VI and VII (2.7.1), 12 isolates sampled in Quebec in 1993/94 were found not to be *C. reinhardtii* (Table S3). To attempt to identify these isolates to the species level, we Sanger sequenced the plastid *rbcL* locus, using the primer pair F1 and R8 for amplification, and the F1, F9, R8 and R10 primers for sequencing (Nozaki et al. 1997; Nozaki et al. 1999). Two isolates (CC-3067 and CC-3081) grew poorly under standard laboratory conditions, and sequencing was not performed for these isolates. Only two isolates (CC-3074, CC-3088) were found to be members of the core-*Reinhardtinia*, the clade containing *C. reinhardtii*, *Volvox carteri*, and their relatives (Nakada et al., 2016). The best megablast hit for CC-3074 had only 95.6% identity to *Pandorina unicocca* (a multicellular species), potentially indicating that this isolate represents a novel unicellular *Reinhardtinia* species, or at least a species for which no *rbcL* sequences have been produced.

2.7.5 Genomic site class annotations

Genomic coordinates for CDS were downloaded for the *C. reinhardtii* genome annotation v5.3 from Phytozome (https://phytozome.jgi.doe.gov/pz/). Within CDS, 0D and 4D sites were defined relative to the reference genome. All "N" bases in the reference genome (~4 Mb) were removed. Any codons that overlapped more than one reading frame, or that contained more than one SNP, were filtered due to the difficulty in determining the degeneracy of sites in such cases.

2.7.6 Population structure analyses

fineSTRUCTURE v2.1.3 was run in "linked" mode, using the flag "-ploidy 1", and otherwise default parameters. Genetic distances between each SNP were calculated assuming a uniform recombination rate, based on the genome-wide estimate of 1.2 x 10^{-5} cM/bp obtained by Liu et al. (2018) from whole-genome re-sequencing of the progeny of crosses between the Quebec isolates CC-2935 and CC-2936. fineSTRUCTURE can be used to probabilistically assign individuals to populations, however we interpreted the results solely based on the coancestry matrix, since fineSTRUCTURE did not cluster isolates effectively into populations. This is likely due to extensive LD and the low number of isolates, resulting in nearly all of the isolates exhibiting a unique relationship to each other in terms of genetic ancestry.

STRUCTURE v2.3.4 was run on a dataset of 4D SNPs subsampled every 20 kb, based on the average decay of LD in *C. reinhardtii* (Flowers et al. 2015). The admixture model with correlated allele frequencies was used, with 20 replications for each value of K (1-10, with the maximum based on the ten sampling locations), a burn-in of 500,000 iterations, and run length of 1,000,000 iterations. Due to the unbalanced sampling (21 isolates from Farnham, Quebec, relative to at most three isolates from any other site), the parameter alpha (that reflects relative admixture between populations) was set as variable for each population, following Wang (2017). STRUCTURE HARVESTER v0.6.94 (Earl and Vonholdt 2012) was used to determine the optimal K using the ΔK method (Evanno et al. 2005). CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to align population assignments across replicates of each K value, using the greedy algorithm with default parameters.

As a complementary approach to visualise multilocus patterns of genetic similarity between isolates, PCA was performed on the 4D SNP dataset used for the STRUCTURE analysis, using the R packages SNPRelate v1.8.0 and gdsfmt v1.10.1 (Zheng et al. 2012). A neighbour joining tree was produced using MEGA v7.0.26 (Kumar et al. 2016) from all 4D sites, using the Tamura-Nei substitution model, and 1000 bootstrap replicates. To test for the presence of isolation by distance within NA1 and NA2, a Mantel test (n=999 permutations) was performed independently for each lineage on a pairwise matrix of 4D genetic distance (calculated using MEGA, Tamura-Nei model) and geographic distance, using vegan v2.4-5 (Oksanen et al. 2017).

2.7.7 Mitochondrial and plastid haplotype networks

Sites that passed filtering were extracted for the entire mitochondrial genome (7.39 kb) and plastid CDS (18.25 kb). PopART (Leigh and Bryant 2015) was used to produce haplotype networks for each organelle using the TCS algorithm (Clement et al. 2002).

2.7.8 Admixture profiling and identification of putatively introgressed genomic regions

Marker SNPs were assigned to each lineage by identifying sites where the within-lineage consensus allele (defined as an allele with ≥ 60% frequency) differed between the two lineages. This resulted in a total of 758,420 marker SNPs, or on average ~135 SNPs per 20 kb. For each isolate, the proportions of marker SNPs matching the NA1 or NA2 consensus were then calculated in 20 kb sliding windows (with 4 kb increments). Intervals of at least five overlapping windows exhibiting a majority of marker SNPs for the alternate lineage to which the isolate belonged were then merged to form putatively introgressed genomic intervals. To visualise the admixture analysis, for each isolate in discrete 20 kb windows the proportions of SNPs with NA1 and NA2 identities were plotted as a heat map along each chromosome.

2.7.9 Identification of genomic tracts inherited identical by descent

We identified genomic tracts that are likely to have been inherited without recombination from a common ancestor (i.e. identical by descent) using the haploid-specific hidden Markov model hmmIBD (Schaffner et al. 2018). This approach infers identical by descent tracts shared between pairs of individuals as genomic regions that are identical by state (allowing for genotyping error), based on SNP allele frequencies, the distance between SNPs in bases, and a genome-wide recombination rate. Additionally, the program estimates the expected proportion of the genome inherited identical by descent between pairs ($\hat{\pi}_{IBD}$) based on the average per-SNP probability of identity by descent, independent of the designation of tracts (Taylor et al. 2017). hmmIBD was run independently for each N. American lineage (NA1/NA2), assuming a recombination rate of 1.2 x 10^{-5} cM/bp (Liu et al. 2018) and otherwise default parameters. As we observed that the majority of identified tracts were within the range of the decay of LD in *C. reinhardtii* (~20 kb), tract length filters of >100 kb (~1.2 cM) and >500 kb (~6.0 cM) were applied. Identical by descent tracts have recently been defined using similar length cut-offs to explore population-level tract sharing (Wakeley and Wilton 2016). Following Carmi et al. (2013), the cohort-averaged sharing was calculated for each isolate as the mean proportion of the genome shared identical by descent between the isolate in question and all other isolates in the sample.

2.7.10 Calculation of population genetics statistics within and between lineages

Genetic diversity was calculated as the average number of pairwise differences per site (π, Nei and Li (1979)) for each of the lineages (NA1/NA2/JPN), and for each sampling site and time point containing two or more isolates. As a measure of differentiation, F_{st} was calculated between each lineage using the approach of Hudson et al. (1992), where within-population π was calculated as an unweighted mean of π for the two lineages in the comparison. As a measure of genetic distance between-lineages*,* we calculated the number of pairwise differences between two random sequences drawn from each lineage (d_{xy}, Nei and Li (1979)).The proportions of fixed, shared and private polymorphisms were calculated for each between lineage comparison. All calculations were performed using custom Perl scripts.

2.8 Acknowledgements

I thank Ahmed Hasan and three anonymous reviewers for their helpful comments on an earlier version of this chapter. The sequencing in this chapter was funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC), the European Research Council (ERC), the National Sciences and Engineering Research Council (NSERC), Canada Foundation for Innovation, and the Yamagata Prefectural Government and Tsuruoka City, Japan.

Chapter 3

Comparative Genomics of *Chlamydomonas*

3.1 Preface

The work in this chapter has been published as a manuscript in *The Plant Cell* and the first-person plural is used throughout to maintain consistency. Minor changes have been made to the published version to preserve formatting across the thesis. I performed all analyses, wrote the first draft of the manuscript and produced all figures and tables with the exception of section 3.4.6 ("Evolution of the mating type locus in *Chlamydomonas*"). The first draft and initial underlying analyses of 3.4.6 were written and performed by Ahmed Hasan, who also produced Figure 6 and Dataset S7. I contributed the analysis on codon usage bias, the associated text and Figure S7 to section 3.4.6. Rob Ness prepared samples and performed DNA extraction for the Illumina sequencing of one of the species. General references to Chapter 2 are cited as Craig et al. (2019).

Citation:

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

3.2 Abstract

Despite its role as a classical model organism in plant sciences, the green alga *Chlamydomonas reinhardtii* entirely lacks genomic resources for any closely related species. We present highly contiguous and well-annotated genome assemblies for three unicellular relatives of the species, *Chlamydomonas incerta*, *Chlamydomonas schloesseri* and the more distantly related *Edaphochlamys debaryana*. The three *Chlamydomonas* genomes are highly syntenous with similar gene contents, although the 129.2 Mb *C. incerta* and 130.2 Mb *C. schloesseri* assemblies are more repeat-rich than the 111.1 Mb *C. reinhardtii* genome. We identify the major centromeric repeat in *C. reinhardtii* as a LINE transposable element homologous to *Zepp* (the centromeric repeat in *Coccomyxa subellipsoidea*) and infer that centromere locations and structure are likely conserved in *C. incerta* and *C. schloesseri*. We report extensive rearrangements, but limited gene turnover, between the *minus* mating type loci of the *Chlamydomonas* species. We produce an 8-species core-*Reinhardtinia* whole-genome alignment, which we use to identify several hundred false positive and missing genes in the *C. reinhardtii* annotation and >260,000 evolutionary conserved elements in the *C. reinhardtii* genome. In summary, these novel resources enable comparative genomics analyses to be performed for *C. reinhardtii*, significantly developing the analytical toolkit for this important model system**.**

3.3 Introduction

With the rapid increase in genome sequencing over the past two decades, comparative genomics analyses have become a fundamental tool in biological research. As the first sets of genomes for closely related eukaryotic species became available, pioneering comparative studies led to refined estimates of gene content and orthology, provided novel insights into the evolution of genome architecture and the extent of genomic synteny between species, and enabled the proportions of genomes evolving under evolutionary constraint to be estimated for the first time (Mouse Genome Sequencing Consortium 2002; Cliften et al. 2003; Stein et al. 2003; Richards et al. 2005). As additional genomes were sequenced it became possible to produce multiple species whole-genome alignments (WGA) and to identify conserved elements (CEs) in noncoding regions for several of the most well-studied lineages (Siepel et al. 2005; Stark et al. 2007; Gerstein et al. 2010; Lindblad-Toh et al. 2011). Many of these conserved noncoding sequences overlap regulatory elements, and the identification of CEs has proved to be among the most accurate approaches for discovering functional genomic sequences (Alföldi and Lindblad-Toh 2013). WGAs are also powerful resources for directly improving gene annotations, with applications including the identification of novel genes, splice forms and exons (Lin et al. 2007; Mudge et al. 2019), distinguishing between protein-coding and long noncoding RNA (lncRNA) genes (Pauli et al. 2012), and the identification of non-standard protein-coding features such as translational frameshifts and stop codon readthrough (Lin et al. 2007; Jungreis et al. 2011).

The ability to perform comparative analyses is contingent on the availability of genome assemblies for species that span a range of appropriate evolutionary distances. While this state has been achieved for most model organisms, there remain several species of high biological significance that entirely lack genomic resources for any closely related species. Hiller et al. (2013) described such cases as 'phylogenetically isolated genomes', specifically referring to species for which the most closely related genomes belong to species divergent by one or more substitutions, on average, per neutrally evolving site. At this scale of divergence an increasingly negligible proportion of the genome can be aligned at the nucleotide-level (Margulies et al. 2006), limiting comparative analyses to the protein-level and impeding the development of such species as model systems in numerous research areas.

Although the ~111 Mb haploid genome of *C. reinhardtii* was among the earliest eukaryotic genomes to be sequenced (Grossman et al. 2003; Merchant et al. 2007), it currently meets the 'phylogenetically isolated' definition. The closest confirmed relatives of *C. reinhardtii* that have genome assemblies belong to the clade of multicellular algae that includes *Volvox carteri*, the *Tetrabaenaceae-Goniaceae-Volvocaceae*, or TGV clade. As introduced in 1.2.2, *C. reinhardtii* and the TGV clade are collectively part of the highly diverse order Volvocales, and the more taxonomically limited clades *Reinhardtinia* and core-*Reinhardtinia* (Nakada et al. 2008; Nakada et al. 2016). Although these species are regularly considered close relatives, multicellularity likely originated in the TGV clade over 200 million years ago (Herron et al.

2009), and *C. reinhardtii* and *V. carteri* are more divergent from one another than human is to chicken (Prochnik et al. 2010).

Without a comparative genomics framework, the wider application of *C. reinhardtii* as a model system is impeded. While this broadly applies to the general functional annotation of the genome as outlined above (e.g. refinement of gene models and annotation of CEs), it is particularly relevant to the field of molecular evolution. Without genomic resources for closely related species it is currently impossible to perform several key analyses, such as the comparison of substitution rates at synonymous and non-synonymous sites of protein-coding genes (i.e. calculating dN/dS), and the inference of ancestral states at polymorphic sites (a requirement of several population and quantitative genetics models (Keightley and Jackson 2018)).

Furthermore, *V. carteri* and the wider TGV clade are extensively used to study the evolution of multicellularity and other major evolutionary transitions (e.g. isogamy to anisogamy), and five genomes of multicellular species spanning a range of organismal complexities have now been assembled (Prochnik et al. 2010; Hanschen et al. 2016; Featherston et al. 2018; Hamaji et al. 2018). These studies have often included analyses of gene family evolution, reporting expansions in families thought to be functionally related to multicellularity. While these analyses have undoubtedly made important contributions, they are nonetheless limited in their phylogenetic robustness, since *C. reinhardtii* is the only unicellular relative within hundreds of millions of years available for comparison. Thus, the availability of annotated genomes for unicellular relatives of *C. reinhardtii* will also serve as an important resource towards reconstructing the ancestral core-*Reinhardtinia* gene content, potentially providing novel insights into the major evolutionary transitions that have occurred in this lineage.

In this chapter we present highly contiguous and well-annotated genome assemblies for the two closest known relatives of *C. reinhardtii*, namely *Chlamydomonas incerta* and *Chlamydomonas schloesseri*, and a more distantly related unicellular species, *Edaphochlamys debaryana*. Via comparison to the genomes of *C. reinhardtii* and the TGV clade species we present the first insights into the comparative genomics of *Chlamydomonas*, focussing specifically on the conservation of genome architecture between species and the landscape of sequence conservation in *C. reinhardtii*. While forming only one of the initial steps in this process, by providing the first comparative genomics framework for the species we anticipate that these novel resources will greatly aid in the continued development of *C. reinhardtii* as a model organism.

3.4 Results and Discussion

3.4.1 The closest known relatives of *Chlamydomonas reinhardtii*

Although the genus *Chlamydomonas* consists of several hundred unicellular species it is highly polyphyletic (Pröschold et al. 2001), and *C. reinhardtii* is more closely related to the multicellular TGV clade than the majority of *Chlamydomonas* species. Given their more conspicuous morphology, the TGV clade contains ~50 described species (Herron et al. 2009), while the unicellular lineage leading to *C. reinhardtii* includes only two other confirmed species, *C. incerta* and *C. schloesseri* (Pröschold et al. 2005; Pröschold et al. 2018). As *C. reinhardtii* is the type species of *Chlamydomonas*, these three species collectively comprise the monophyletic genus (Figure 1A, B, C), and throughout this chapter *Chlamydomonas* will be used specifically to refer to this clade.

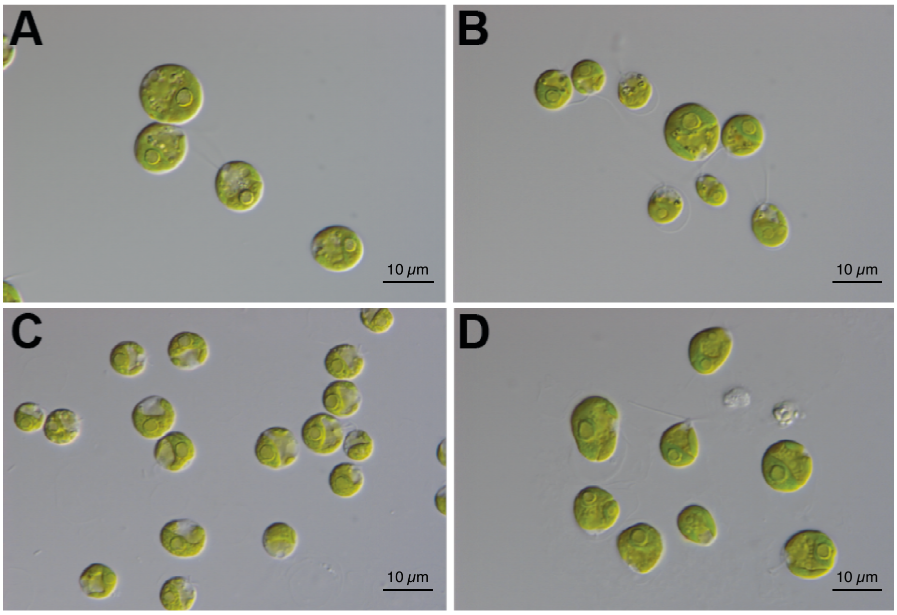


Figure 1. Images of *Chlamydomonas* and *Edaphochlamys* species.
(A) *C. reinhardtii*.
(B) *C. incerta* SAG 7.73
(C) *C. schloesseri* SAG 2486 (=CCAP 11/173).
(D) *E. debaryana* SAG 11.73 (=CCAP 11/70).
All images provided by Thomas Pröschold.

C. incerta is the closest known relative of *C. reinhardtii*, and a small number of comparative genetics analyses have been performed between the two species (Ferris et al. 1997; Popescu et al. 2006; Smith and Lee 2008). *C. incerta* is known from only two isolates, and we selected the original isolate SAG 7.73 for sequencing. Unfortunately, although *C. incerta* SAG 7.73 is nominally from Cuba, the geographic origin of the isolate is uncertain due to a proposed historical culture replacement with *C. globosa* SAG 81.72 from the Netherlands (Harris et al. 1991). As the direction of replacement is unknown, the strain may be from either location. SAG 7.73 is currently listed as *C. globosa* based on the taxonomic reassessment of Nakada et al. (2010), although Pröschold and Darienko (2018) contested this change. We refer to SAG 7.73 as *C. incerta* given its existing use in the genetics literature. *C. schloesseri* was recently described by Pröschold et al. (2018), with three isolates from a single site in Kenya in culture. We selected CCAP 11/173 for sequencing.

Beyond *Chlamydomonas* there are a substantial number of unicellular core-*Reinhardtinia* species with uncertain phylogenetic relationships (i.e. that may be part of the lineage including *Chlamydomonas*, the lineage including the TGV clade, or outgroups to both). Among these, the best studied is *E. debaryana*, which was recently renamed from *Chlamydomonas debaryana* (Pröschold et al. 2018). *E. debaryana* appears to be highly abundant in nature (unlike the three *Chlamydomonas* species), with more than 20 isolates from across the Northern Hemisphere in culture, suggesting that it could be developed as a model for studying algal molecular ecology. Draft genomes of the isolates NIES-2212 from Japan (Hirashima et al. 2016) and WS7 from the USA (Nelson et al. 2019) were recently assembled, while we selected CCAP 11/70 from the Czech Republic for sequencing (Figure 1D).

3.4.2 The genomes of *Chlamydomonas incerta*, *Chlamydomonas schloesseri* and *Edaphochlamys debaryana*

Using a combination of Pacific Biosciences (PacBio) sequencing for *de novo* assembly (40-49x coverage, Table S1) and Illumina sequencing for error correction (43-86x coverage, Table S2), we produced contig-level genome assemblies for *C. incerta*, *C. schloesseri* and *E. debaryana*. All three assemblies were highly contiguous, with N50s of 1.6 Mb (*C. incerta*), 1.2 Mb (*C. schloesseri*) and 0.73 Mb (*E. debaryana*), and L50s of 24, 30 and 56 contigs, respectively (Table 1). Genome-mode BUSCO (Benchmarking Universal Single-Copy Ortholog) scores (Waterhouse et al. 2018) supported a high-level of assembly completeness, with the percentage of universal chlorophyte single-copy orthologs identified in each genome ranging from 95.9% to 98.1%. These metrics compare favourably to the best existing core-*Reinhardtinia* (Table 1) and Volvocales assemblies (Dataset S2). Although the *C. reinhardtii* and *V. carteri* assemblies have greater scaffold-level N50s than the three new assemblies, they are both considerably more fragmented at the contig level, with N50s of 215 kb and 85 kb, respectively. While this is not surprising given our application of long read sequencing, it nonetheless demonstrates that these important model genomes could be substantially improved by additional sequencing effort. The contig-level N50s of the three new assemblies also exceeded those of recent TGV clade assemblies, namely *Gonium pectorale* (Hanschen et al. 2016) and the PacBio-based assemblies of *Yamagishiella unicocca* and *Eudorina* sp. 2016-703-Eu-15 (hereafter *Eudorina* sp.) (Hamaji et al. 2018).

Assembled genome size varied moderately across the eight species, ranging from 111.1 Mb (*C. reinhardtii*) to 184.0 Mb (*Eudorina* sp.) (Table 1). Both *C. incerta* (129.2 Mb) and *C. schloesseri* (130.2 Mb) had consistently larger assemblies than *C. reinhardtii*, and the *E. debaryana* assembly (142.1 Mb) was larger than those of *Y. unicocca* and *V. carteri*. Although additional genome assemblies or flow cytometry estimates will be required to fully explore genome size evolution in the core-*Reinhardtinia*, these results suggest that *C. reinhardtii* may have undergone a recent reduction in genome size. Furthermore, while earlier comparisons between multicellular species and *C. reinhardtii* led to the observation that certain metrics of genomic complexity (e.g. gene density and intron length, see below) correlate with organismal complexity, these results indicate that genome size, at least for

Table 1. Genome assembly metrics for eight high-quality core-*Reinhardtinia* genome assemblies.

Species	*Chlamydomonas reinhardtii* v5	*Chlamydomonas incerta*	*Chlamydomonas schloesseri*	*Edaphochlamys debaryana*	*Gonium pectorale*	*Yamagishiella unicocca*	*Eudorina. sp. 2016-703-Eu-15*	*Volvox carteri* v2
Assembly level	chromosome	contig	contig	contig	scaffold	contig	scaffold	scaffold
Assembly size (Mb)	111.10	129.24	130.20	142.14	148.81	134.23	184.03	131.16
Number of contigs/scaffolds	17*	453	457	527	2373	1461	3180	434
N50 (Mb)	7.78	1.58	1.21	0.73	1.27	0.67	0.56	2.60
Contig N50 (Mb)	0.22	1.58	1.21	0.73	0.02	0.67	0.30	0.09
L50	7	24	30	56	30	53	83	15
Contig L50	141	24	30	56	1871	53	155	410
GC (%)	64.1	66.0	64.4	67.1	64.5	61.0	61.4	56.1
TEs & satellites (Mb / %)	15.33 / 13.80	26.75 / 20.70	27.48 / 21.11	20.05 / 14.11	11.65 / 7.83	29.57 / 22.03	46.81 / 25.43	22.22 / 16.94
Simple & low complexity repeats (Mb / %)	8.71 / 7.84	8.57 / 7.72	10.19 / 9.17	6.40 / 5.76	4.15 / 3.74	6.55 / 4.88	15.15 / 8.23	6.45 / 5.80
BUSCO genome mode (complete % / fragmented %)	96.5 / 1.7	96.5 / 1.6	96.1 / 1.7	94.0 / 1.9	86.3 / 4.5	95.9 / 2.2	94.7 / 2.7	95.9 / 2.4

*17 chromosomes + 37 unplaced scaffolds
BUSCO was run using the Chlorophyta odb10 dataset. See Dataset S2 for complete BUSCO results.

these species, does not. Conversely, as proposed by Hanschen et al. (2016), GC content does appear to decrease with increasing cell number, with genome-wide values ranging from 64.1 to 67.1% for the unicellular species and from 64.5 to 56.1% in the TGV clade (Table 1).

The larger genome sizes of the unicellular species, relative to *C. reinhardtii,* can largely be attributed to differences in the content of transposable elements (TEs) and satellite DNA (defined as tandem repeats with monomers >10 bp). We produced repeat libraries for each species by combining manual curation (Dataset S1) with automated repeat identification. For *C. reinhardtii,* we produced an exhaustively curated library that updates all sequences in the existing library available from Repbase (https://www.girinst.org/repbase/) and more than doubles the total number of annotated TEs (269 vs 119 subfamilies), details of which are presented in 5.4.1. For the three new assemblies, we performed targeted curation of the most abundant TEs in each species, similar to the annotation performed for the *V. carteri* genome project (Prochnik et al. 2010). All three of the new assemblies contained greater total amounts (20.1-27.5 Mb) and higher genomic proportions (14.1-21.1%) of complex repetitive sequence than *C. reinhardtii* (15.3 Mb and 13.8%) (Table 1). As discussed below, the larger genome size of *E. debaryana* can also be partly attributed to the substantially higher number of genes present in the species. For all three assemblies, repeat content was relatively consistent across contigs, except for small contigs (<~100 kb), which exhibited highly variable repeat contents and likely represent fragments of complex regions that have resisted assembly (Figure S1). The higher repeat contents of the three assemblies were broadly consistent across TE orders (Figure S2), although a direct comparison of the TEs present in each genome is complicated by phylogenetic bias. The inclusion of a curated repeat library for *C. reinhardtii* directly contributes to masking and repeat classification in related species, however this effect will become increasingly negligible as divergence increases. This is likely to at least partly explain the lower repeat content and higher proportion of "unknown" classifications observed for *E. debaryana* relative to *C. incerta* and *C. schloesseri* (Table 1, Figure S2).

Nonetheless, based on the manual curation of the most abundant TE families, a qualitative comparison is possible. All curated TEs belonged to orders and superfamilies that are present in one or both of *C. reinhardtii* and *V. carteri*, suggesting a largely common repertoire of TEs across the core-*Reinhardtinia*. Alongside more widely recognised elements such as *L1* LINEs and *Gypsy* LTRs, all species contained families of the comparatively obscure *Dualen* LINE elements (Kojima and Fujiwara 2005), *PAT*-like DIRS elements (Poulter and Butler 2015) and *Helitron2* rolling-circle elements (Bao and Jurka 2013). We also identified *Zisupton* and *Kyakuja* DNA transposons, both of which were reported as potentially present in *C. reinhardtii* upon their recent discovery (Böhne et al. 2012; Iyer et al. 2014). Although not the focus of this study, the annotation of elements from such understudied superfamilies highlights the importance of performing manual TE curation in phylogenetically diverse lineages. Alongside improving our understanding of TE biology, these elements are expected to contribute towards more effective repeat masking/classification and gene model annotation in related species, which will be of increasing importance given the large number of chlorophyte genome projects currently in progress (Blaby-Haas and Merchant 2019).

3.4.3 Phylogenomics of the core-*Reinhardtinia* and Volvocales

Due to the low number of available genomes and gene annotations, the phylogenetics of the Volvocales has almost exclusively been studied using ribosomal and plastid marker genes. These analyses have successfully delineated several broad clades (e.g. *Reinhardtinia*, *Moewusinia*, *Dunaliellinia*) (Nakada et al. 2008), but often yielded inconsistent topologies for more closely related taxa. Utilising both our own and several recently published genomic resources, we further explored the phylogenomic structure of the core-*Reinhardtinia* and Volvocales. As several genomes currently lack gene annotations, we first used an annotation-free approach based on the identification of chlorophyte single-copy orthologs with BUSCO. This dataset consisted of 1,624 genes, present in at least 15 of the 18 included species (12 *Reinhardtinia*, three other Volvocales, and three outgroups from the Sphaeropleales, Dataset S2). For the 11 species with gene annotations (Dataset S3), we produced a second dataset based on the orthology clustering of each species' proteome, which yielded 1,681 single-copy orthologs shared by all species. For both datasets, we performed maximum-likelihood (ML) analyses using IQ-TREE (Nguyen et al. 2015). Analyses were performed on both concatenated protein alignments (producing a species-tree) and individual alignments of each ortholog (producing gene trees), which were then summarised as a species-tree using ASTRAL-III (Zhang et al. 2018).

All four of the resulting phylogenies exhibited entirely congruent topologies, with near maximal support values at all nodes (Figure 2, Figure S3). Rooting the tree on the Sphaeropleales species, the monophyly of the Volvocales, *Reinhardtinia* and core-*Reinhardtinia* clades were recovered. *Chlamydomonas* was recovered with the expected branching order (Pröschold et al. 2018), as was the monophyly and expected topology of the TGV clade (Nakada et al. 2019). The most contentious phylogenetic relationships are those of the remaining unicellular core-*Reinhardtinia*, which include *E. debaryana* and the recently published genomes of *Chlamydomonas sphaeroides* (Hirashima et al. 2016) and *Chlamydomonas* sp. 3112 (Nelson et al. 2019). In the most gene-rich analysis to date, *E. debaryana* grouped in a weakly-supported clade with *Chlamydomonas* (termed metaclade C), while *C. sphaeroides* grouped with a small number of other unicellular species on the lineage including the TGV clade (Nakada et al. 2019). In our analysis, *E. debaryana* and *C. sphaeroides* were recovered as sister taxa on the lineage including *Chlamydomonas*, meeting the prior definition of metaclade C as the sister clade of the TGV clade and its unicellular relatives. Due to its recent discovery, *Chlamydomonas* sp. 3112 has not been included in previous phylogenetic analyses. We classified *Chlamydomonas* sp. 3112 as a member of the core-*Reinhardtinia* based on sequence similarity of ribosomal and plastid genes, which suggested that it is likely a close relative of *Chlamydomonas zebra* (Table S3). Given its basal phylogenetic position relative to metaclade C and the TGV clade, species such as *Chlamydomonas* sp. 3112 could prove particularly useful in future efforts to reconstruct the ancestral gene content of the core-*Reinhardtinia*.

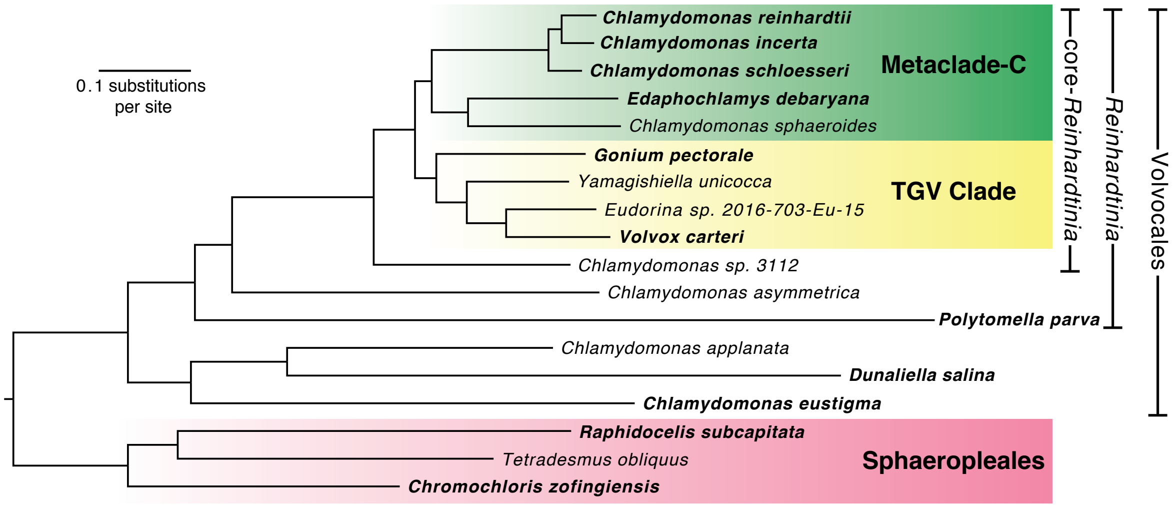


Figure 2. Maximum likelihood phylogeny of 15 Volvocales species and three outgroups. The phylogeny was inferred using the LG+F+R6 model and a concatenated protein alignment of 1,624 chlorophyte BUSCO genes. All ultrafast bootstrap values ≥99%. Species in bold have gene model annotations and were included in the OrthoFinder-based phylogenies (Figures S3B, C). Phylogeny was rooted on the three Sphaeropleales species (highlighted in pink).

3.4.4 Conserved genome architecture and centromeric structure in *Chlamydomonas*

Almost nothing is known about karyotype evolution and the rate of chromosomal rearrangements in *Chlamydomonas* and the core-*Reinhardtinia*. Prochnik et al. (2010) reported that the syntenic genomic segments identified between *C. reinhardtii* and *V. carteri* contained fewer genes than human and chicken syntenic segments, in part due to a greater number of small inversions disrupting synteny. As the longest contigs in our assemblies were equivalent in length to *C. reinhardtii* chromosome arms (6.4, 4.5 and 4.2 Mb for *C. incerta*, *C. schloesseri* and *E. debaryana*, respectively), we explored patterns of synteny between the three species and *C. reinhardtii*. We used SynChro (Drillon et al. 2014) to identify syntenic segments, which first uses protein sequence reciprocal best-hits to anchor syntenic segments, before extending segments via the inclusion of homologs that are syntenic but not reciprocal best-hits. All three *Chlamydomonas* genomes were highly syntenous, with 99.5 Mb (89.5%) of the *C. reinhardtii* genome linked to 315 syntenic segments spanning 108.1 Mb (83.6%) of the *C. incerta* genome, and 98.5 Mb (88.6%) of the *C. reinhardtii* genome linked to 409 syntenic segments spanning 108.1 Mb (83.1%) of the *C. schloesseri* genome.

Given the high degree of synteny, it was possible to order and orientate the contigs of *C. incerta* and *C. schloesseri* relative to the assembled chromosomes of *C. reinhardtii* (Figure 3). A substantial proportion of the *C. reinhardtii* karyotype appeared to be conserved in *C. incerta*, with six of the 17 chromosomes (1, 3, 4, 7, 14 and 16) showing no evidence of inter-chromosomal rearrangements, and a further three (5, 13 and 15) showing evidence for only minor translocations <150 kb in length (Figure 3A). Consistent with its greater divergence from *C. reinhardtii*, *C. schloesseri* exhibited such one-to-one conservation between only four chromosomes (5, 7, 11 and 14) (Figure 3B). For both species, patterns of synteny indicated at least one inter-chromosomal rearrangement affecting the remaining chromosomes, although without additional scaffolding of contigs it is difficult to comment on the effect of such rearrangements on karyotype. Furthermore, by direct comparison to *C. reinhardtii* chromosomes we may have overestimated karyotype conservation due to undetected chromosome fusion/fission events (i.e. if a *C. reinhardtii* chromosome is present as two chromosomes in one of the related species). For both *C. incerta* and *C. schloesseri*, all chromosomes (except chromosome 15 in the *C. incerta* comparison) contained intra-chromosomal rearrangements relative to *C. reinhardtii*, most of which were small inversions spanning <100 kb (Figure S4A, B). Synteny was far weaker between *C. reinhardtii* and *E. debaryana*, with 58.6 Mb (52.8%) of the *C. reinhardtii* genome linked to 1,975 syntenic segments spanning 64.8 Mb (45.6%) of the *E. debaryana* genome (Figure S4C). Taken together with the previous assessment of synteny between *C. reinhardtii* and *V. carteri*, these results suggest that karyotype evolution in the core-*Reinhardtinia* is expected to be dynamic, with generally high levels of synteny but a non-negligible rate of inter-chromosomal rearrangements present between closely related species, and likely far greater karyotypic diversity present between more distantly related species.

Given the high-contiguity and synteny of the assemblies, it was possible to assess features of genome architecture that regularly resist assembly in short-read assemblies. Telomeric repeats were observed in all three assemblies, with six *C. incerta* and 19 *C. schloesseri* contigs terminating in the sequence (TTTTAGGG)$_n$, and 15 *E. debaryana* contigs terminating in (TTTAGGG)$_n$ (Dataset S4). The *Arabidopsis*-type sequence (TTTAGGG)$_n$ is ancestral to green algae and was previously confirmed as the telomeric repeat in *E. debaryana*, while the derived *Chlamydomonas*-type sequence (TTTTAGGG)$_n$ is found in both *C. reinhardtii* and *V. carteri* (Fulnečková et al. 2012). Given the phylogenetic relationships in Figure 2, this implies either two independent transitions to the derived sequence or a reversion to the ancestral sequence in the lineage including *E. debaryana*, providing further evidence for the relatively frequent transitions that have produced extensive variation in telomere composition in green algae and land plants (Peska and Garcia 2020). Ribosomal DNA repeats (rDNA) were assembled as part of three larger contigs in both *C. incerta* and *C. schloesseri*, but were found only as fragmented contigs entirely consisting of rDNA in *E. debaryana*. Although poorly assembled in *C. reinhardtii*, the rDNA arrays are located at subtelomeric locations on chromosomes 1, 8 and 14, where cumulatively they are estimated to be present in 250-400 tandem copies (Howell 1972; Marco and Rochaix 1980). The assembled *C. incerta* and *C. schloesseri* rDNA arrays (which are not complete and are present in five tandem copies at most) were entirely syntenous with those of *C. reinhardtii*, suggesting conservation of subtelomeric rDNA organisation in *Chlamydomonas* (Figure 3). The architecture and evolution of subtelomeric regions in *C. reinhardtii* and the three new assemblies presented here have recently been described by Chaux-Jukic et al. (2021).

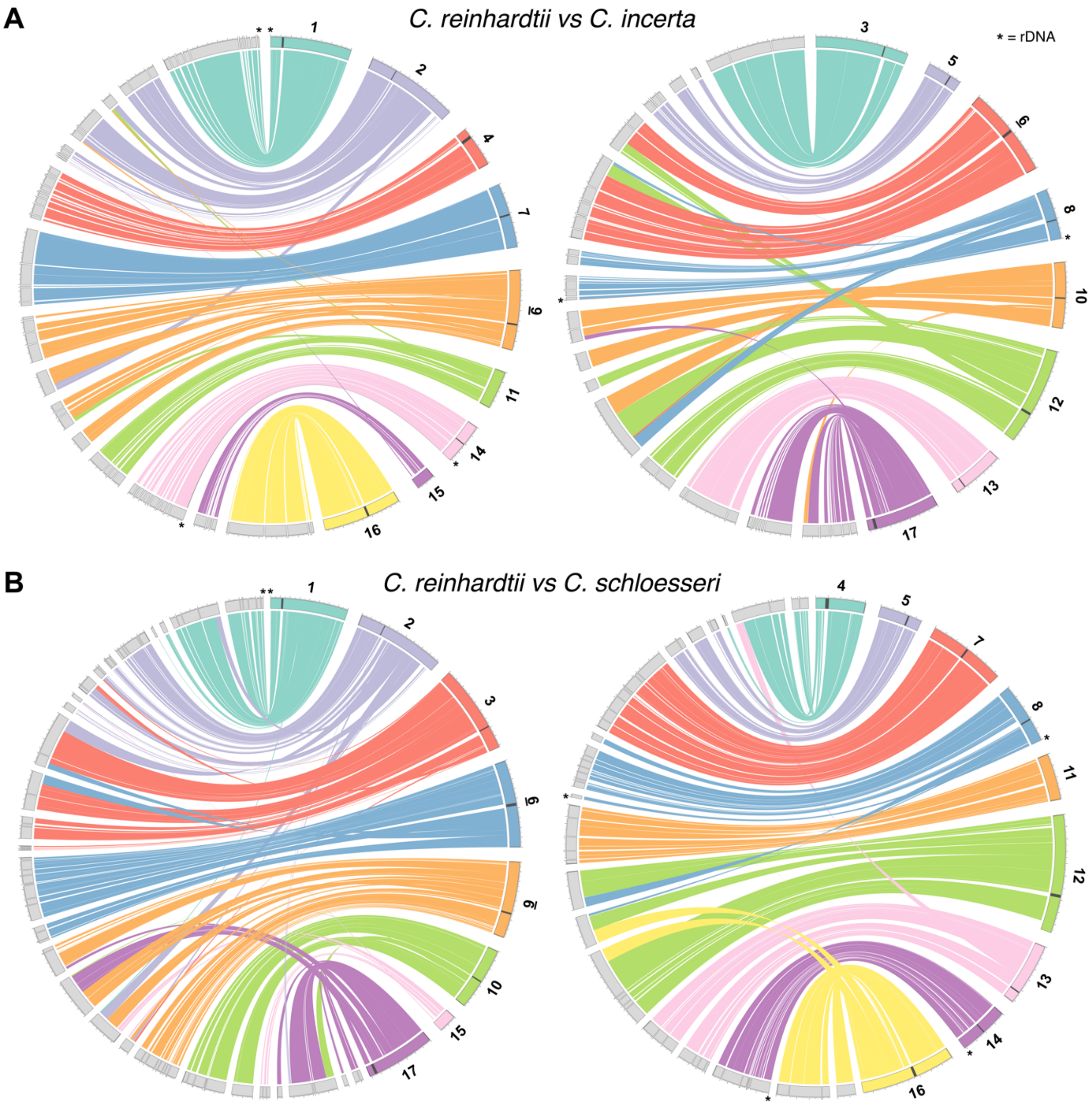


Figure 3. Synteny between *C. reinhardtii* and it's close relatives.
Circos plots (Krzywinski et al. 2009) between *C. reinhardtii* and *C. incerta* **(A)** and *C. reinhardtii* and *C. schloesseri* **(B)**. *C. reinhardtii* chromosomes are represented as coloured segments and split across the left and right plots in each panel, and *C. incerta* / *C. schloesseri* contigs are shown as grey segments. Contigs are arranged and orientated relative to *C. reinhardtii* chromosomes, and adjacent contigs with no signature of rearrangement are plotted without gaps. Dark grey bands highlight putative *C. reinhardtii* centromeres and asterisks represent rDNA. Not that the colours representing specific chromosomes differ between the panels.

Finally, we were able to assess the composition and potential synteny of centromeres in *Chlamydomonas*. The centromeric locations of 15 of the 17 *C. reinhardtii* chromosomes were recently mapped by Lin et al. (2018), who observed that these regions were characterised by multiple copies of genes encoding reverse transcriptase. Upon inspection of these regions, we found that the majority of these genes are encoded by copies of the *L1* LINE element *L1*-

1_CR (Kapitonov and Jurka 2004a). Although these regions are currently not well-enough assembled to conclusively define the structure of centromeric repeats, *L1-1_CR* is present in multiple copies at all 15 putative centromeres and appears to be the major centromeric component (with chromosome-specific contributions from other TEs, especially *Dualen* LINE elements) (Table S4, Figure S5A). Remarkably, phylogenetic analysis of all curated *L1* elements from green algae indicated that *L1-1_CR* is more closely related to the *Zepp* elements of *Coccomyxa subellipsoidea* than to any other *L1* elements annotated in *C. reinhardtii* (Figure 4A). The divergence of the classes Trebouxiophyceae (to which *C. subellipsoidea* belongs) and Chlorophyceae (to which *C. reinhardtii* belongs) occurred in the early Neoproterozoic era (i.e. 700-1,000 million years ago) (Del Cortona et al. 2020), implying that *L1-1_CR* has been evolving independently from all other *C. reinhardtii L1* elements for more than half a billion years. *Zepp* elements are thought to constitute the centromeres in *C. subellipsoidea*, where they are strictly present as one cluster per chromosome (Blanc et al. 2012). The clustering pattern of *Zepp* arises due to a nested insertion mechanism that targets existing copies, creating tandem arrays consisting mostly of the 3' end of the elements (due to frequent 5' truncations upon insertion) (Higashiyama et al. 1997). Chromosome-specific clustering of *L1-1_CR* was also evident in *C. reinhardtii*, with highly localised clusters observed at all 15 of the putative centromeres (Figure 4B). The double-peaks in *L1-1_CR* density present on chromosomes 2, 3 and 8, and the single sub-telomeric cluster present on chromosome 5, are all the result of misassemblies in these highly repetitive regions (Chapter 4). Thus, outside the putative centromeres, *L1-1_CR* appears to be entirely absent from the *C. reinhardtii* genome. To distinguish the updated annotation of *L1-1_CR* in our repeat library from the original Repbase version, we propose the name *ZeppL-1_cRei*, where *ZeppL* represents *Zepp*-like.

Every putative centromeric location in *C. reinhardtii* coincided with breaks in syntenic segments and the termination of contigs in *C. incerta* and *C. schloesseri* (Figure 3), suggesting that these regions are also likely to be repetitive in both species. The phylogenetic analysis revealed the presence of one and two *ZeppL-1_cRei* homologs in *C. incerta* and *C. schloesseri,* respectively (Figure 4A). Of the 30 contig ends associated with the 15 *C. reinhardtii* centromeres, 28 contigs in both species contained a *ZeppL* element within their final 20 kb (Figure S5B, C), and genome-wide the *ZeppL* elements exhibited similarly localised clustering to that observed in *C. reinhardtii* (Figure S6A, B). Thus, it appears that both the location and composition of the *C. reinhardtii* centromeres are likely conserved in *C. incerta* and *C. schloesseri*. We identified two families of *ZeppL* elements in the *E. debaryana* genome and one family of *ZeppL* elements in the *Eudorina* sp. genome, although we did not find any evidence for *ZeppL* elements in either *Y. unicocca* or *V. carteri*. Given the lack of synteny between *C. reinhardtii* and *E. debaryana* it was not possible to assign putatively centromeric contigs. Nonetheless, highly localised genomic clustering of *ZeppL* elements was observed for both *E. debaryana* and *Eudorina* sp. (Figure S6C, D), suggesting that these elements may play a similar role to that in *Chlamydomonas*.

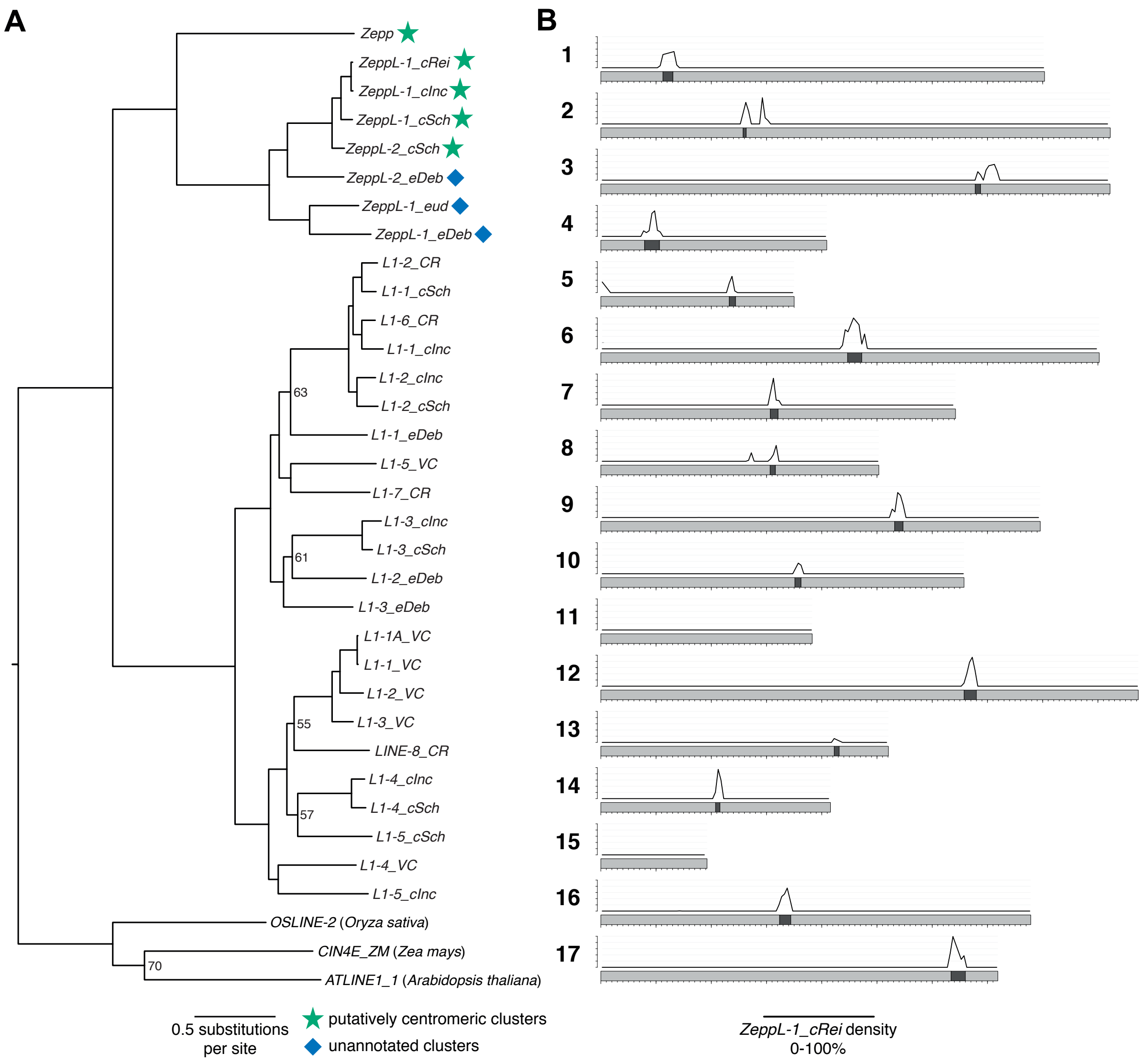


Figure 4. Phylogenetic relationship and centromeric clustering of *Zepp*-like elements.
(A) Maximum likelihood phylogeny of chlorophyte *L1* elements inferred using the LG+F+R6 model and alignment of endonuclease and reverse transcriptase domains. Bootstrap values ≤70% are shown. Phylogeny was rooted on three plant *L1* elements. Species are provided by suffixes: *CR*/*cRei* = *C. reinhardtii*; *VC* = *V. carteri*; *cInc* = *C. incerta*; *cSch* = *C. schloesseri*; *eDeb* = *E. debaryana*; *eud* = *Eudorina* sp. 2016-703-Eu-15.
(B) Density (0-100%) of *ZeppL-1_cRei* in 50 kb windows across *C. reinhardtii* chromosomes. Dark bands represent putative centromeres, x-axis ticks represent 100 kb increments and y-axis ticks 20% increments. Plot produced using karyoploteR (Gel and Serra 2017). Note that *ZeppL-1_cRei* is a synonym of the Repbase element *L1-1_CR*.

Centromeres, the chromosomal regions at which kinetochores assemble during cell division, exhibit substantial structural diversity across eukaryotes. In most species a specific nucleosome, centromere-specific histone H3 (cenH3, also known as CENP-A or CENH3, amongst others), is localised at the centromere and is required for kinetochore assembly (Talbert and Henikoff 2020). At one extreme, point centromeres are determined by short specific sequences, such as an ~125 bp sequence in *Saccharomyces cerevisiae* that features a single cenH3 nucleosome (Furuyama and Biggins 2007). Many species have short regional centromeres several kb in length, for example in *Plasmodium falciparum* cenH3 is localised to 4-4.5 kb regions centred on AT-rich sequences (Hoeijmakers et al. 2012). At the other extreme, in many plant and animal genomes centromeres are defined by complex arrays of satellite DNA that can reach lengths of ~1 Mb, although cenH3 is not present at all satellite monomers (Talbert and Henikoff 2020). Differing altogether from regional centromeres, holocentromeres are located at multiple locations along the length of "holocentric" chromosomes (Mandrioli and Manicardi 2020). Generally falling somewhere between short regional centromeres and satellite centromeres in length (i.e. 10s to 100s of kb), transposon-rich centromeres feature TEs as their main structural constituent. These centromeres can be primarily based on a single TE family or on several different TEs without any apparent organisation. For example, in the slime mold *Dictyostelium discoideum* the DIRS retrotransposon *DIRS-1* constitutes ~50% of centromeric sequence and is absent from the rest of the genome (Glöckner and Heidel 2009), while in *Neurospora crassa* the centromeres are defined by AT-rich regions enriched for various degenerated TEs (Smith et al. 2011). Other recent examples revealed by long-read sequencing include *Drosophila melanogaster*, where centromeres were previously thought to consist of satellite DNA but were shown to coincide with regions featuring a specific LINE element *G2/Jockey-3* (Chang et al. 2019), and the oomycete *Phytophthora sojae*, where centromeres were defined by a *Copia* LTR element *CoLT* (Fang et al. 2020). In both cases, homologous TEs were found at the putative centromeres of closely related species.

It therefore seems likely that centromeres in *Chlamydomonas* can be categorised as transposon-rich, and primarily based on a specific TE in the *ZeppL* elements. The lengths and repeat content of the putative *C. reinhardtii* centromeres are revisited in greater detail in 4.4.3 using long-read based assemblies. Given the evolutionary distance between *C. subellipsoidea* and *Chlamydomonas,* it is tempting to predict that *ZeppL* elements may be present at the centromeres of many other species of green algae. However, it is unlikely that centromeres are conserved between species from the Trebouxiophyceae and Chlorophyceae. First, centromeric repeats in the Chlorophyceae species *Chromochloris zofingiensis* consist of entirely unrelated *Copia* LTR elements (Roth et al. 2017). Second, the apparent absence of *ZeppL* elements from *Y. unicocca* and *V. carteri* suggest that these elements are not required for centromere formation in these species. Instead, it is possible that the propensity for *Zepp* and *ZeppL* elements to form clusters may play a role in their recruitment as centromeric sequences, which is likely to have happened independently in *C. subellipsoidea* and *Chlamydomonas*. As more highly contiguous chlorophyte assemblies become available, it will be important to search these genomes for *ZeppL* clusters to assess whether these elements can be used more generally as centromeric markers.

3.4.5 Gene and gene family evolution in the core-*Reinhardtinia*

We performed gene annotation for each species using 7.4-8.2 Gb of stranded RNA-seq (Table S5). Protein mode BUSCO scores supported a high level of annotation completeness across all three species (97.0-98.1% chlorophyte genes present), although relative to genome mode scores there was an increase in the proportion of fragmented genes (4.0-5.9%) (Table

2). *C. incerta* and *C. schloesseri* had comparable gene counts to *C. reinhardtii*, although lower gene densities due to their larger genomes. With 19,228 genes, the *E. debaryana* genome contained substantially more genes than any other currently annotated core-*Reinhardtinia* species. As reported by Hanschen et al. (2016), several metrics appeared to correlate with organismal complexity. Relative to the unicellular species, gene density was lower, and median intergenic and intronic lengths were longer, in *G. pectorale* and *V. carteri*. Presumably this is at least partly due to an increase in the amount of regulatory sequence in these genomes, although this has not yet been explored.

Table 2. Gene annotation metrics for core-*Reinhardtinia* species.

Species	*C. reinhardtii* v5.6*	*C. incerta*	*C. schloesseri*	*E. debaryana*	*G. pectorale*	*V. carteri* v2.1
Number of genes	16,656	16,350	15,571	19,228	16,290	14,247
Number of transcripts	18,311	16,957	16,268	20,450	16,290	16,075
Gene coverage (Mb / %)	91.22 / 82.10	94.42 / 73.06	94.29 / 73.42	103.13 / 72.55	65.04 / 43.71	84.00 / 64.04
UTR coverage (Mb / %)	17.32 / 15.59	14.51 / 11.22	12.02 / 9.23	14.68 / 9.31	0 / 0	15.15 / 11.55
Mean intron number	7.81	8.58	7.67	9.31	6.15	6.73
Median intron length (bp)	229	225	244	198	310	343
Median intergenic distance (bp)	134	341	408	555	2372	905
BUSCO protein mode (complete % / fragmented %)	96.1 / 2.3	91.1 / 5.9	94.7 / 3.0	94.1 / 4.0	81.5 / 12.9	94.7 / 2.0

**C. reinhardtii* metrics are based on a customised repeat-filtered version of v5.6 (3.6.8, 4.4.8). Intron metrics are based only on introns present within coding sequence to avoid differences caused by variation in the quality of UTR annotation. BUSCO was run using the Chlorophyta odb10 dataset, see Dataset S3 for complete BUSCO results.

Across all species, both mean intron lengths (3.4.9) and intron numbers per gene were very high for such compact genomes. For the unicellular species, the mean number of introns present in coding sequence (CDS) per gene ranged from 7.7-9.3, with slightly lower counts in *G. pectorale* (6.2) and *V. carteri* (6.7). These numbers are more comparable to vertebrates such as human (8.5) than to other model organisms with similar genomes sizes, such as *Caenorhabditis elegans* (5.1), *D. melanogaster* (3.0), and *Arabidopsis thaliana* (4.1). Modelling of intron evolution across the breadth of eukaryota has predicted that a major expansion of introns occurred early in chlorophyte evolution, and that high intron densities have since been maintained in certain lineages by a balance between intron loss and gain (Csuros et al. 2011). It has been hypothesised that the relative roles of DNA double-strand break repair pathways play a major role in the dynamics of intron evolution, as homologous recombination (HR) is thought to cause intron deletion, while non-homologous end-joining (NHEJ) may result in both intron gain and loss (Farlow et al. 2011). HR occurs at an extremely low rate in *C. reinhardtii* (Zorin et al. 2005), and if this is shared across the core-*Reinhardtinia* it may contribute to the maintenance of such high intron numbers. Alternatively, introns could be maintained by other forces, such as selection. Interestingly, high rates of NHEJ have recently been linked to high GC content in prokaryotes (Weissman et al. 2019), and it may be the case that double strand break repair is generally an important and underappreciated force in *Chlamydomonas* genome evolution.

To explore gene family evolution in the core-*Reinhardtinia,* we performed orthology clustering using the six available high-quality gene annotations (98,342 total protein-coding genes), which resulted in the delineation of 13,728 orthogroups containing 86,446 genes (Figure 5). The majority of orthogroups (8,532) were shared by all species, with the second most abundant category (excluding genes unique to a single species) being those present in all species except *G. pectorale* (868 orthogroups). Given the lower BUSCO score observed for *G. pectorale* (Table 2) it is likely that a proportion of these orthogroups are also universal to core-*Reinhardtinia* species. The next most abundant category was the 859 orthogroups present only in *Chlamydomonas*. Unfortunately, essentially nothing is known about the biology and ecology of *C. incerta* and *C. schloesseri*, and even for *C. reinhardtii* we have a minimal understanding of its biology in natural environments (Sasso et al. 2018; Craig et al. 2019). Nonetheless, more than 30% of the *Chlamydomonas*-specific orthogroups were associated with at least one functional domain (Dataset S5). The most common association was with protein kinase domains (50 orthogroups), followed by other relatively common domains in *C. reinhardtii* including peptidase M11/gametolysin (14 orthogroups). *C. reinhardtii* is known to encode a large kinome relative to other unicellular green algae (Wheeler et al. 2008), with 575 *C. reinhardtii* genes annotated with protein kinase domains in our current analysis, 86 of which were present in *Chlamydomonas*-specific orthogroups. With 51 genes, the most gene-rich *Chlamydomonas*-specific orthogroup represented the *NCL* (nuclear control of chloroplast gene expression-like) gene family. These genes encode RNA binding proteins of unknown function, are entirely absent from *V. carteri*, and are undergoing a rapid diversification in *C. reinhardtii* via recurrent gene duplication that has formed a cluster of at least 32 genes on chromosome 15 (Boulouis et al. 2015). Both *C. incerta* and *C. schloesseri* contained six genes in the *NCL* orthogroup, all of which are syntenous with chromosome 15 in *C. reinhardtii*. It therefore appears that although the *NCL* genes evolved in the common ancestor of *Chlamydomonas*, most of the diversification is specific to *C. reinhardtii* itself and attempts to uncover the evolutionary driver of the rapid expansion could focus on biological differences between *C. reinhardtii* and its closest relatives. In contrast to *Chlamydomonas*, only 51 orthogroups were unique to the two multicellular species. This may be an underestimate due to the relative incompleteness of the *G. pectorale* annotation, and it will be important to re-visit this analysis as more annotations become available. Nonetheless, the availability of the three new high-quality annotations for unicellular species will provide a strong comparative framework to explore the relative roles of gene family birth versus expansions in existing gene families in the transition to multicellularity.

Finally, we explored the contribution of gene family expansions to the high gene count of *E. debaryana*. The *E. debaryana* genome contained more species-specific genes (3,556) than any other species, however this figure was not substantially higher than the unassigned gene counts for *G. pectorale* and *V. carteri* (Figure 5). We quantified *E. debaryana* gene family expansion and contraction by calculating per orthogroup log2-transformed ratios of the *E. debaryana* gene count and the mean gene count for the other species. Arbitrarily defining an

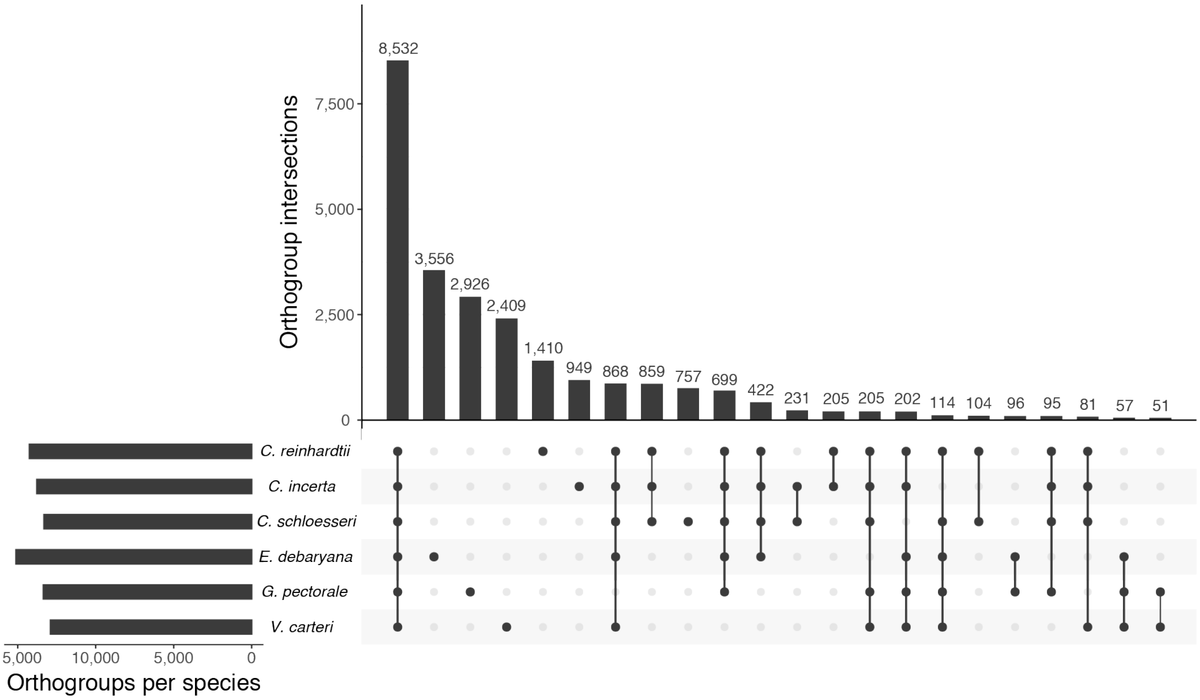


Figure 5. Gene families in the core-*Reinhardtinia*.
Upset plot (Lex et al. 2014) representing the intersection of orthogroups between species. Numbers above the bars represent the number of orthogroups shared by a given intersection. Only intersections with at least 50 orthogroups are shown.

expansion as a log2-transformed ratio >1 (i.e. a given orthogroup containing more than twice as many *E. debaryana* genes than the mean of the other species) and a contraction as a ratio <-1, we identified *E. debaryana*-specific expansions in 294 orthogroups and contractions in 112. With 16 genes in *E. debaryana* relative to at most one in the other five species, the most expanded orthogroup contained genes encoding scavenger receptor cysteine-rich (SRCR) and C-type lectin (CTL) domains (Dataset S6). SRCR and CTL domains have roles in innate immunity in animals and the presence of >30 genes encoding SRCR and/or CTL domains in *C. reinhardtii*, which may have roles in immunity or other processes such as chemoreception, was a surprising finding from the genome project (Wheeler et al. 2008). These large gene families have also been shown to have variable gene copy number among isolates of *C. reinhardtii* (Flowers et al. 2015). Other orthogroups exhibiting the most extreme expansions were associated with HIT and MYND-type zinc fingers, polyketide cyclase SnoaL-like domains, protein kinase domains and pherophorins (Dataset S6), although in all cases the *C. reinhardtii* and *V. carteri* genes present in these orthogroups were not annotated with specific functions. Furthermore, more than 100 of the expanded orthogroups were not associated with any functional domains at all. Only ~50% of *C. reinhardtii* genes are annotated with domains and only ~10% are formally annotated with primary gene symbols (Blaby and Blaby-Haas 2017). Further exploring the relationships between gene content and the biological differences of *C. reinhardtii* and its close relatives may be a powerful approach to functionally characterise additional genes, especially those that are unique to specific clades such as the Volvocales or core-*Reinhardtinia*.

3.4.6 Evolution of the mating type locus in *Chlamydomonas*

Across core-*Reinhardtinia* species, sex is determined by a haploid mating-type locus with two alleles, termed *plus* (MT^{+}) or female, and *minus* (MT^{-}) or male, in isogamous and anisogamous species. The *C. reinhardtii* mating type locus is located on chromosome 6, spanning >400 kb and consisting of three domains, the T (telomere-proximal), R (rearranged) and C (centromere-proximal) domains. While both the T and C domains exhibit high synteny between the mating type alleles, the crossover-suppressed R domain contains the only mating type-specific genes (Ferris and Goodenough 1997) and harbours substantial structural variation, featuring several inversions and rearrangements (Ferris et al. 2002; De Hoff et al. 2013). As detailed in 1.3.6, comparative analyses of MT^{+}/female and MT^{-}/male haplotypes between *C. reinhardtii* and TGV clade species have revealed highly dynamic evolution, with extensive gene turnover and structural variation resulting in a complex and discontinuous evolutionary history of haplotype reformation (Ferris et al. 2010; Hamaji et al. 2016b; Hamaji et al. 2018). Only one mating type-specific gene is common to all species, the minus dominance gene (*MID*), which determines MT^{-}/male gametic differentiation (Ferris and Goodenough 1997).

To explore whether mating type evolution is similarly dynamic between the more closely related *Chlamydomonas* species, we used a reciprocal best-hit approach to identify *C. reinhardtii* orthologs in *C. incerta* and *C. schloesseri*. The sequenced isolates of both species were inferred to be MT^{-} based on the presence of *MID*, as was previously reported for *C. incerta* (Ferris et al. 1997). Orthologs of *MTD1*, the second and only other MT^{-}-limited gene in *C. reinhardtii*, were also identified in both species. Although we were able to map the entire *C. reinhardtii* MT^{-} haplotype to single contigs in both the *C. incerta* and *C. schloesseri* assemblies, it is important to state that it is currently impossible to define the R domain boundaries for either species without sequencing their MT^{+} alleles. Unfortunately, it is currently unknown if any of the one (*C. incerta*) or two (*C. schloesseri*) other isolates are MT^{+}, and as no isolate from either species has been successfully crossed it is not even known if they are sexually viable (Pröschold et al. 2005). Furthermore, as sexual reproduction has not been observed for either species it cannot definitively be stated that they are heterothallic, as *MID* orthologs are present and required for sexual development in homothallic species in the TGV clade (Hamaji et al. 2013; Yamamoto et al. 2017a). To test this possibility, we explored patterns of synonymous codon usage in both species. Assuming patterns of recombination are similar to those in *C. reinhardtii*, if *C. incerta* and *C. schloesseri* are heterothallic it would be expected that *MID* (and possibly also *MTD1*) would exhibit little evidence of selection acting on codon usage, due to low selection efficacy caused by the absence of recombination (both crossovers and gene conversion). Indeed, *MID* in *C. incerta* was previously shown to have the lowest codon adaptation index (CAI) amongst a dataset of 67 genes (Popescu et al. 2006). We quantified codon adaptation for all genes using the index of translation elongation (I_{TE}), a metric that accounts for mutation bias (unlike CAI) but can otherwise be interpreted analogously (Xia 2015). In both species, *MID* was within the lowest 2% of genes for I_{TE} genome-wide and had the lowest I_{TE} of any genes present on the contigs syntenous to the *C. reinhardtii* mating type (Figure S7). *MTD1* also exhibited low I_{TE} in *C.*

schloesseri (lowest ~9% of genes), although the reduction in *C. incerta* was less pronounced (lowest ~23%). These results support heterothallism in both species, although it is possible that *MTD1* may not be *MT*⁻-specific in *C. incerta* (as is found in *Y. unicocca* and *Eudorina* sp. (Hamaji et al. 2018)). We therefore proceed with this assumption, although it will be a priority to confirm this via sequencing of the other existing isolates or new isolates in the future. Finally, we also determined the sequenced isolate of *E. debaryana* to be *MT*⁻ via the identification of *MID*, although we did not explore mating type evolution further given the evolutionary distance to *C. reinhardtii*. Unlike *C. incerta and C. schloesseri*, heterothallic mating pairs of *E. debaryana* are in culture, and a future comprehensive study of the mating type locus in the species is therefore possible.

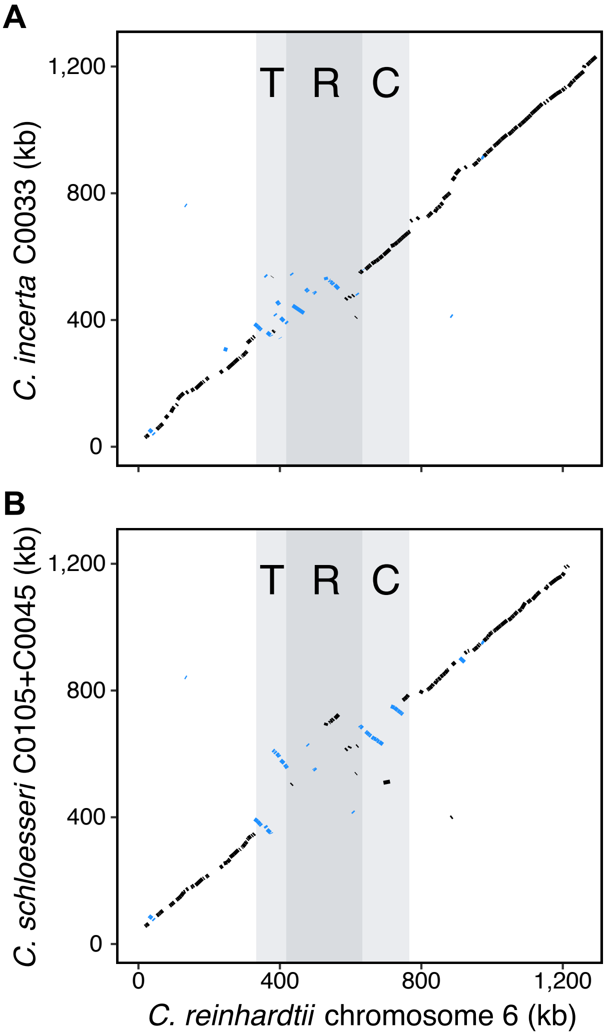


Figure 6. Synteny across the *C. reinhardtii MT*⁻ haplotype and inferred *MT*⁻ haplotypes in *C. incerta* and *C. schloesseri*.
Each line represents an individual *C. reinhardtii* gene and its inferred ortholog in **(A)** *C. incerta* and **(B)** *C. schloesseri*. The T, R and C domains of the *C. reinhardtii MT*⁻ haplotype are highlighted. Genes with inverted orientations are shown in blue. For *C. incerta* the entire genomic region plotted was syntenic to contig C0033. For *C. schloesseri*, the entire *MT*⁻ haplotype was syntenic to C0045, but C0105 was appended to show synteny extending to the most telomere-proximal region of *C. reinhardtii* chromosome 6.

In *C. incerta,* gene order was entirely syntenic across the C domain, except for *MT0828*, which did not yield a hit anywhere in the genome. Conversely, both T and R domain genes have undergone several rearrangements and inversions relative to *C. reinhardtii MT*$^{–}$ (Figure 6A). Furthermore, the T domain genes *SPP3* and *HDH1* were present on separate contigs in *C. incerta* and do not appear to be mating type-linked (Dataset S7). Synteny otherwise continued well into the adjacent autosomal sequence, in line with the genome-wide patterns of synteny (3.4.3). We observed even less synteny between *C. reinhardtii* and *C. schloesseri MT*$^{–}$ genes, with both the T and C domains showing two large inversions each (Figure 6B). However, gene order in the surrounding autosomal sequence was also largely collinear. As in *C. incerta*, *SPP3* was located elsewhere in the *C. schloesseri* assembly, suggesting a relatively recent translocation to the T domain in *C. reinhardtii*. The T domain gene *97782* was also located on a different contig, while the genes *MT0796, MT0828* and *182389* did not yield hits anywhere in the *C. schloesseri* genome. Finally, we found no hits for the *MT*$^{+}$-specific genes *FUS1* and *MTA1* in either species, suggesting that these genes (assuming they exist) are also expected to be *MT*$^{+}$-specific in *C. incerta* and *C. schloesseri*.

The lack of collinearity relative to the *C. reinhardtii* T domain may be indicative of an extended R domain in these species, especially in *C. schloesseri*, where we observe multiple rearrangements in all three domains. We do not, however, observe dramatic variation in mating type size; whereas *C. reinhardtii MT*$^{–}$ is ~422 kb, if *NIC7* and *MAT3* are taken as the boundaries of the locus (De Hoff et al. 2013), *C. incerta MT*$^{–}$ is ~329 kb and *C. schloesseri MT*$^{–}$ is ~438 kb. In all, while we do find evidence of *MT*$^{–}$ haplotype reformation within *Chlamydomonas*, this is mostly limited to rearrangements, with far less gene turnover and locus size variation than has been observed between more distantly related core-*Reinhardtinia* species. While mating type evolution has previously been explored in the context of transitions from unicellularity to multicellularity and isogamy to anisogamy, our data suggest that mating type haplotype reformation is still expected to occur between closely related isogamous species, albeit at a reduced scale.

3.4.7 Alignability and estimation of neutral divergence

In order to facilitate the identification of conserved elements (CEs) and an assessment of the current *C. reinhardtii* gene models, we produced an 8-species core-*Reinhardtinia* WGA using Cactus (Armstrong et al. 2019). Based on the alignment of *C. reinhardtii* four-fold degenerate (4D) sites extracted from the WGA, we estimated putatively neutral branch lengths across the topology connecting the eight species under the GTR substitution model (Figure 7A). Divergence between *C. reinhardtii* and *C. incerta,* and *C. reinhardtii* and *C. schloesseri*, was estimated as 34% and 45%, respectively. Divergence between *C. reinhardtii* and *E. debaryana* was estimated as 98%, while all four TGV clade species were saturated relative to *C. reinhardtii* (i.e. on average, each 4D site is expected to have experienced more than one substitution). To put these estimates within a more recognisable framework, divergence across *Chlamydomonas* is approximately on the scale of human-rodent divergence (Lindblad-Toh et al. 2011), while divergence between *Chlamydomonas* and the TGV clade is roughly equivalent to that of mammals and sauropsids (birds and reptiles), which diverged ~320 million years ago (Alföldi et al. 2011). Our estimates corroborate a previous estimate of synonymous divergence between *C. reinhardtii* and *C. incerta* of 37% (Popescu et al. 2006) and are broadly in line with the divergence time estimate of ~230 million years between the TGV clade and their unicellular ancestors (Herron et al. 2009). However, it is important to note that we have likely underestimated neutral divergence, as 4D sites are unlikely to be evolving neutrally due to selection acting on codon usage, which has been shown to reduce divergence between *C. reinhardtii* and *C. incerta* (Popescu et al. 2006).

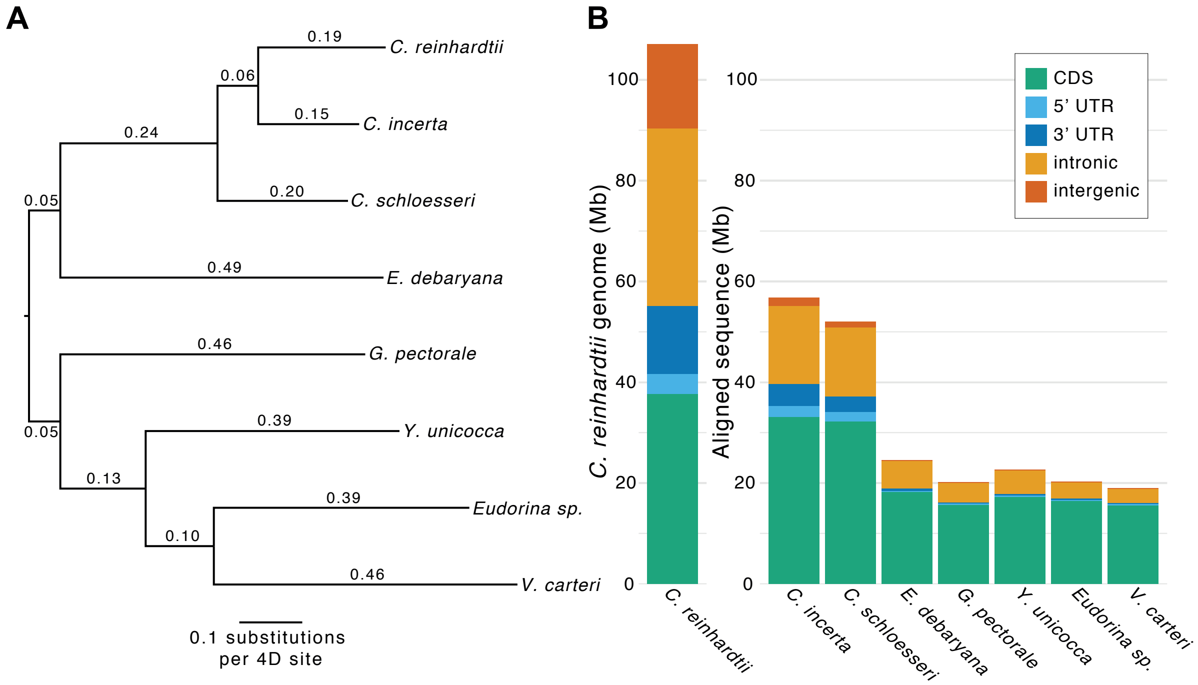


Figure 7. Putatively neutral divergence and genome-wide alignability across the core-*Reinhardtinia*.
(A) Estimates of putatively neutral divergence under the GTR model, based on the topology of Figure 2 and 1,552,562 4D sites extracted from the Cactus WGA.
(B) A representation of the *C. reinhardtii* genome by site class, and the number of aligned sites per *C. reinhardtii* site class for each other species in the Cactus WGA.

As expected, genome-wide alignability (the proportion of bases aligned between *C. reinhardtii* and a given species in the WGA) decreased substantially with increasing divergence, with 53.0% of the *C. reinhardtii* genome aligned to *C. incerta*, 48.6% to *C. schloesseri*, and on average only 19.9% to the remaining five species (Figure 7B). The majority of *C. reinhardtii* CDS was alignable within *Chlamydomonas* (87.7% and 85.5% to *C. incerta* and *C. schloesseri*, respectively), indicating that it will be possible to perform molecular evolutionary analyses (e.g. calculating dN/dS) between the three species. CDS also constituted most of the aligned sequence to the other five species, comprising on average 78.3% of the aligned bases despite forming only 35.2% of the *C. reinhardtii* genome. In contrast, far less non-exonic sequence was alignable, especially beyond *Chlamydomonas*. Substantial proportions of intronic bases were aligned to *C. incerta* (44.1%) and *C. schloesseri* (38.8%), with on average 11.3% aligned to the other five species. Less than 10% of intergenic sequence was aligned to any one species, and on average less than 1% was aligned to non-*Chlamydomonas* species. Distributions of intergenic tract lengths across the core-*Reinhardtinia* are highly skewed (Figure S8), so that in *C. reinhardtii* tracts shorter than 250 bp constitute 63.5% of tracts but just 5.5% of total intergenic sequence. The sequence content of tracts >250 bp is highly repetitive (total repeat content 63.4%), while tracts <250 bp are relatively free of repeats (4.3% repeat content) and as a result are far more alignable to *C. incerta* and *C. schloesseri* (40.8% and 32.0% of bases aligned, respectively). This suggests that at least for introns and short intergenic tracts it is feasible to explore the landscape of non-exonic evolutionary constraint, primarily utilising alignment data from *Chlamydomonas,* supplemented by what is expected to be alignment of only the most conserved sites at greater evolutionary distances.

3.4.8 False positive and missing genes in the *Chlamydomonas reinhardtii* v5.6 annotation

One of the major successes of comparative genomics has been the refinement of gene annotations. Many approaches that utilise WGAs rely on the ability to distinguish between protein coding and non-coding sequence, and programs such as PhyloCSF (Lin et al. 2011) quantify coding potential by assessing candidate alignments for evolutionary signatures characteristic of CDS, such as higher synonymous and lower nonsynonymous divergence. Using our new resources, we first attempted to assess the prevalence of false positive genes in the current *C. reinhardtii* v5.6 annotation. Prior to this, we filtered 1,085 genes from the v5.6 annotation (~6% of genes) that appear to be part of TEs (4.4.8). We divided the remaining 16,656 genes into a 'control' set that contained all genes with at least one core-*Reinhardtinia* ortholog and/or a functional domain (15,365 genes), and a 'test' set that failed both conditions (1,291 genes). We ran PhyloCSF on alignments of CDS extracted from the WGA, producing a per gene score (with more positive scores indicating a higher coding potential). The score distributions for the control and test gene sets were strikingly different, with a median score of 359.9 for the control set and 0 for the test set (with scores of 0 in almost all cases representing a complete lack of alignment) (Figure 8A). In full, 865 test set genes (~67%) scored <1, while the same was true for 598 control set genes (~4%). The positive scores and likely true positive status for approximately one third of the test set could be explained by the orthologs of these genes being absent from the annotations for the aligned species (as PhyloCSF is not reliant on gene annotations from outgroup species). Alternatively, many of these genes may be fast evolving at the protein-level, thus escaping orthology clustering. Of the remaining test set genes, caution must be taken in designating false positive status since this subset may include genes unique to *C. reinhardtii* (i.e. orphan genes or recent gene duplications). There is also expected to be a false positive rate associated with PhyloCSF caused by misalignment or a lack of power (i.e. for genes where CDS does not align across several of the species in the WGA), as demonstrated by the ~4% of genes scoring <1 in the control set.

We therefore performed two further analyses to more accurately delineate a set of false positive gene models. First, for each gene we calculated the ratio of genetic diversity (π) at zero-fold degenerate (0D) and 4D sites based on whole-genome re-sequencing data from 17

C. reinhardtii field isolates from Quebec (Craig et al. 2019). As would be expected under an assumption of purifying selection, median $\pi_{0D/4D}$ for the control set was 0.230 and <2% of genes had a ratio >1 (Figure 8B). Conversely, median $\pi_{0D/4D}$ for the test set was 0.665 and ~30% of genes had a ratio >1. Taking the 95th percentile of control $\pi_{0D/4D}$ (0.717) as a cut-off, 823 test set genes exceeded this threshold (or $\pi_{0D/4D}$ could not be calculated at all), 626 of which also had a PhyloCSF score <1. Second, we quantified codon adaptation for each gene using I_{TE}, under the assumption that false positive genes would be expected to deviate from the overall codon usage bias of *C. reinhardtii*. Median I_{TE} for the control set was 0.683, dropping to 0.619 for the test set (Figure 8C). Taking the 5th percentile of control I_{TE} (0.588) as a cut-off, 430 test set genes were below this threshold, 345 of which had a phyloCSF score <1. Considering the three analyses together, 250 test set genes (~19%) had a PhyloCSF score <1 and had $\pi_{0D/4D}$ and I_{TE} values exceeding the control set thresholds, while 721 (~56%) genes had a PhyloCSF score <1 and exceeded one but not both thresholds. We designate these genes as low coding potential, with the exact number of spurious gene models in the v5.6 annotation likely falling somewhere between the sets of 250 and 721 genes.

There are several biological reasons why genuine protein-coding genes may have outlying values in the above analyses. For example, genes evolving under positive selection (e.g. immune system genes) may exhibit an excess of nonsynonymous substitutions or variants, affecting both the PhyloCSF score and $\pi_{0D/4D}$. As with the case of *MID* (3.4.6), genes evolving in low recombination regions may be expected to have sub-optimal codon usage. Nonetheless, there are several additional features of the low coding potential genes that support their likely status as false positive models. Focussing on the set of 250 genes, their open reading frames (ORFs) were considerably shorter (mean 372.2 bp) and consisted of fewer exons (mean 2.1 exons) than the remaining genes (means 2293.2 bp and 8.9 exons). GC content at 3rd codon positions was substantially lower (mean 65.5%) relative to the remaining genes (mean 81.9%), and was only marginally higher than the genome-wide GC content (64.1%) that would be expected in random sequence. Genetic diversity of high impact sites (start codons, 0D sites in stop codons, and splice junctions) was an order of magnitude higher (0.0177) relative to the remaining genes (0.000983) and was of the same order as genetic diversity genome-wide (3.4.9), indicating that many of the ORFs of the low coding potential gene set are disrupted by variants at the population-level. Finally, the putative start codons of low coding potential genes generally lacked strong Kozak sequences, suggesting that they possess unfavourable sequence context for translational initiation. Following Cross (2015), we calculated a ‘Kozak’ score for each gene based on the agreement between the *C. reinhardtii* Kozak consensus sequence and the information content in bits per site for the five bases up and downstream of each start codon. The distribution of Kozak scores for the low coding potential genes more closely resembled random sequence (Figure 8D) and did not produce a recognisable Kozak consensus sequence (Figure S9).

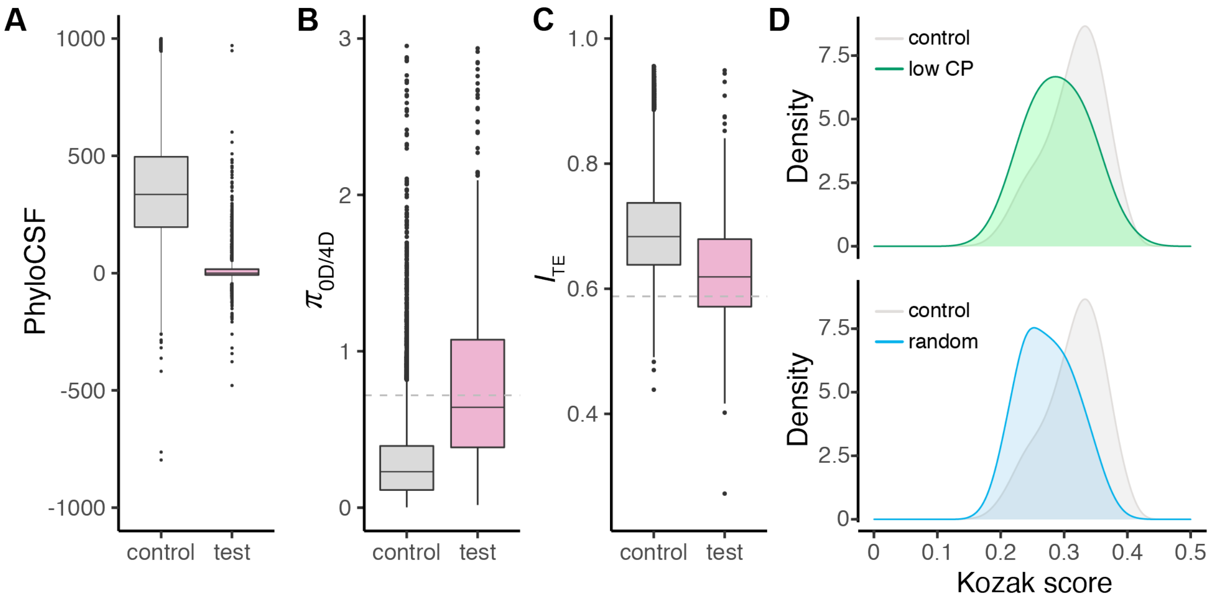


Figure 8. Coding potential analyses and false positive genes.
(A) Boxplot of PhlyoCSF scores for the control and test set genes.
(B) Boxplot of the ratio of genetic diversity at 0D and 4D sites ($\pi_{0D/4D}$) for control and test set genes. Grey dashed line represents 95th percentile of control gene values.
(C) Boxplot of codon adaptation, as quantified by I_{TE} for control and test set genes. Grey dashed line represents 5th percentile of control gene values.
(D) Density plot of 'Kozak scores', quantified as the per gene agreement of the start codon sequence context to that of the *C. reinhardtii* Kozak consensus sequence. Low CP refers to 'low coding potential', and specifically the 250 test set genes that failed all three coding potential analyses. Control refers to the opposite half of the control set genes from that which was used to generate the Kozak consensus sequence (3.6.11). Random was calculated from 10,000 randomly generated sequences based on an average GC content of 64.1%.

Given the complexity and probabilistic nature of gene prediction, the presence of several hundred likely false positives is not unexpected, with even the most developed annotations such as human containing a non-negligible number of dubious gene models (Abascal et al. 2018). This is especially true given the high GC content of *C. reinhardtii,* since the length of ORFs expected by chance increases with GC content as a result of decreasing stop codon frequency (Pohl et al. 2012). The mean ORF length of the low coding potential set (~124 codons) is not substantially longer than the 100 codons that is often used as a statistically robust threshold. Indeed, as there are genuine protein coding genes of <100 amino acids and several functionally characterised lncRNAs that contain spurious ORFs longer >100 codons, a clean designation of coding and noncoding sequence based on ORF length is not possible in any case (Housman and Ulitsky 2016). Assuming that they are expressed, it is possible that many of these gene models are in fact lncRNAs, which have not yet been thoroughly characterised in *C. reinhardtii*. The one study that annotated lncRNAs in the species filtered any transcripts that overlapped existing gene annotations (Li et al. 2016), which despite being a logical approach may have resulted in many lncRNAs being discarded. Given the compactness of the *C. reinhardtii* genome, an alternative possibility is that many of the false positive genes are in fact spurious ORFs within the untranslated regions (UTRs) of neighbouring genes. Further approaches such as long read RNA sequencing will be required to distinguish between such hypotheses.

Finally, we attempted to identify genes missing from v5.6 using a similar comparative approach. We performed *de novo* gene prediction, which yielded 433 novel gene models. We reduced this to 142 high-confidence genes based on the models having either a PhyloCSF score >100 or a syntenic homolog in one or both of *C. incerta* and *C. schloesseri* (based on the SynChro approach, 3.6.11). Supporting their validity, 37 or the 142 genes contained a functional domain. Furthermore, 35 had significant blastp hits (>95% sequence similarity, >=80% query protein length) to *C. reinhardtii* proteins from annotation v4.3 (Dataset S8) and likely represent models that were lost during the transition from v4 to v5 of the genome. This is a known issue with the current annotation, and our re-discovered gene set includes fundamental genes such as *psbW* that have been previously recorded as missing (Blaby and Blaby-Haas 2017). Most interestingly, we recently showed that 25 of these missing genes were part of polycistronic transcripts together with existing genes in the v5.6 annotation (Gallaher et al. 2021). These genes were most likely overlooked by previous annotation tools due to their non-canonical organisation.

3.4.9 The Genomic landscape of sequence conservation in *Chlamydomonas reinhardtii*

Based on the WGA, we identified 265,006 CEs spanning 33.8 Mb or 31.5% of the *C. reinhardtii* genome. The majority of CE sites overlapped CDS (70.6%), with the remaining sites overlapping 5’ UTRs (2.9%), 3’ UTRs (4.4%), introns (20.0%) and intergenic sites (2.0%) (Table 3). Relative to the site class categories themselves, 63.1% of CDS, 24.8% of 5’ UTRs, 11.0% of 3’ UTRs, and 19.2% of intronic sites were overlapped by CEs. Only 4.1% of intergenic sites were overlapped by CEs, however when splitting intergenic tracts into those <250 bp (short tracts) and >250 bp (long tracts), a more appreciable proportion of short tract sites (14.1%) were overlapped by CEs. As would be predicted given the expectation that CEs contain functional sequences, genetic diversity was 39.5% lower for CEs (0.0134) than non-CE bases (0.0220), a result that was relatively consistent across site classes except for long intergenic tracts (Table 3). It is important to state that the identified CEs contain a proportion of non-constrained sites. While this is always to be expected to some extent (e.g. CDS is generally included in CEs despite the presence of synonymous sites), given a mean length of 128 bp our CE dataset should be cautiously interpreted as regions containing elevated proportions of constrained sites.

Given the compactness of the *C. reinhardtii* genome (82.1% genic, median intergenic tract length 134 bp), it is expected that a high proportion of regulatory sequence will be concentrated in UTRs and intergenic sequences immediately upstream of genes (i.e. promoter regions). Relatively little is known about the genome-wide distribution of regulatory elements in *C. reinhardtii*, although analyses based on motif modelling have identified putative *cis*-regulatory elements in these regions (Castruita et al. 2011; Ding et al. 2012; Hamaji et al. 2016a). Presumably many CEs overlapping UTRs and promoter regions harbour regulatory elements, and the CEs we have identified could be used in future studies to validate potential functional motifs (i.e. by assessing whether predicted motifs are overlapped by CEs). However, since the CE lengths are generally considerably longer than the expected length of regulatory elements, genomes for additional close relatives of *C. reinhardtii* (assuming such species exist) would be required to achieve sufficient power to directly identify novel regulatory elements.

Table 3. Overlap between conserved elements and *C. reinhardtii* genomic site classes.

Site class	CE overlap (Mb)	Proportion of CE bases (%)	Proportion of site class (%)	Genetic diversity all sites (π)	Genetic diversity CE sites (π)	Genetic diversity non-CE sites (π)
CDS	23.85	70.64	63.10	0.0144	0.0112	0.0204
5' UTR	0.97	2.86	24.76	0.0189	0.0138	0.0208
3' UTR	1.48	4.38	10.97	0.0205	0.0151	0.0213
intronic	6.76	20.01	19.15	0.0248	0.0216	0.0256
intergenic <250 bp	0.13	0.38	14.07	0.0229	0.0194	0.0235
intergenic ≥250 bp	0.56	1.65	3.55	0.0137	0.0134	0.0138

All six annotated core-*Reinhardtinia* species contained conspicuously long introns (median lengths 198-343 bp, Table 2). As reported previously for *C. reinhardtii* (Merchant et al. 2007), the distribution of intron lengths for core-*Reinhardtinia* species lacked the typical peak in intron lengths at 60-110 bp that is present in several model organisms with similarly compact genomes (Figure 9A, B). In *D. melanogaster*, short introns (<80 bp) appear to largely consist of neutrally evolving sequence, while longer introns that form the tail of the length distribution contain sequences evolving under evolutionary constraint (Halligan and Keightley 2006). To explore the relationship between intron length and sequence conservation in *C. reinhardtii*, we ordered introns by length and divided them into 50 bins, so that each bin contained an approximately equal number (~2,667) of introns. Mean intron length per bin was significantly negatively correlated with the proportion of sites overlapped by CEs (Pearson's r = -0.626, p <0.01) (Figure 9C). This was particularly pronounced for introns <100 bp (~5% of introns), for which 48.1% of sites were overlapped by CEs, compared to 18.5% for longer introns. Therefore, it appears that in a reverse of the situation found in *D. melanogaster*, the minority of introns in *C. reinhardtii* are short and contain a high proportion of conserved sites, while most introns are longer and are expected to contain a higher proportion of sites evolving under little constraint. The tight peak in the distribution of intron lengths combined with the lack of sequence constraint in *D. melanogaster* short introns led Halligan and Keightley (2006) to hypothesise that intron length was under selection, but not the intronic sequence itself, and that introns had essentially evolved to be as short as possible. It is possible that *C. reinhardtii* introns are similarly evolving under selection to be bounded within certain length constraints, although the selective advantage of maintaining intron lengths substantially longer than the minimum remains unknown. Given that atypical intron length distributions are common to all core-*Reinhardtinia* species, whatever mechanism is driving intron length is likely evolutionarily ancient.

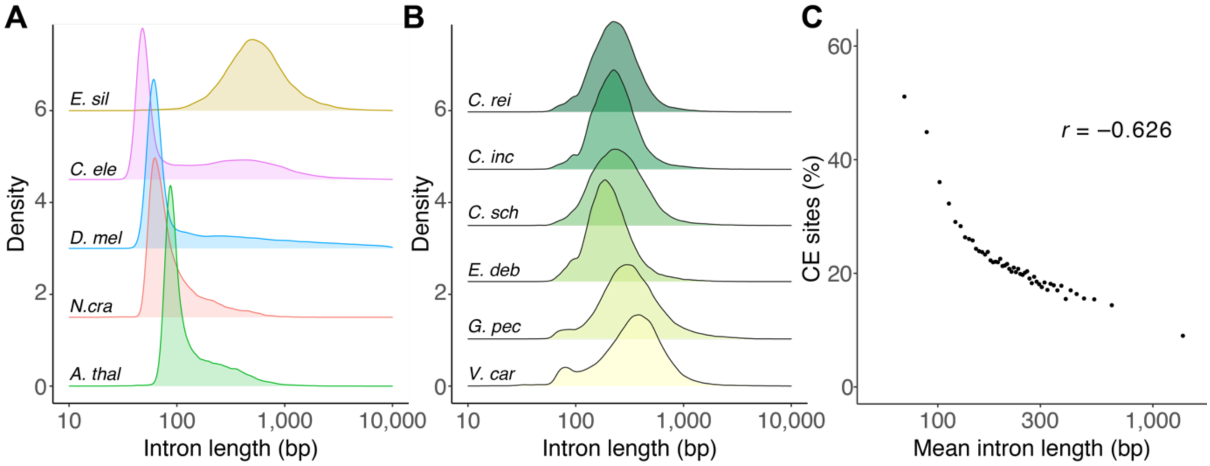


Figure 9. Intron lengths and overlap with conserved elements.
(A) Intron length distributions for five model organisms (*A. thal* = *A. thaliana*, *N. cra* = *Neurospora crassa*, *D. mel* = *D. melanogaster*, *C. ele* = *C. elegans*, *E. sil* = *Ectocarpus siliculosus*). The brown alga *E. siliculosus* is included as an example of an atypical distribution like that found in the core-*Reinhardtinia*.
(B) Intron length distributions for six core-*Reinhardtinia* species (*C. rei* = *C. reinhardtii*, *C. inc* = *C. incerta*, *C. sch* = *C. schloesseri*, *E. deb* = *E. debaryana*, *G. pec* = *G. pectorale*, *V. car* = *V. carteri*).
(C) Correlation between mean intron length per bin and the proportion of sites overlapped by CEs. Introns were ordered by length and separated into 50 bins containing an approximately equal number of introns.

There are several reasons why intronic sites could be evolving under evolutionary constraint. First, alternative splicing (AS) can result in either the entire intron (i.e. intron retention, IR) or part of an intron (alternative acceptor or donor splice sites) being incorporated into mature mRNA. IR is the most common form of AS in *C. reinhardtii* (~30% of events) and occurs significantly more frequently in shorter genes (median = 181 bp) (Raj-Kumar et al. 2017). However, AS in the species has not yet been extensively characterised and only ~1% of introns are currently annotated as alternatively retained. Second, many RNA genes have been identified within introns of protein-coding genes (Chen et al. 2008; Valli et al. 2016). Third, many introns are expected to contain regulatory sequences. This is especially true for introns within the first 1 kb, which for many genes have strong regulatory effects on gene expression (Rose 2018). The addition of a specific first intron to transgenes in *C. reinhardtii* has been shown to substantially increases their expression (Baier et al. 2018), and introns closer to the 5’ end of genes appear to tolerate fewer TE insertions in the species (Philippsen et al. 2016). Short introns <100 bp represented the first intron in a gene approximately four-fold more frequently (44.6%) than longer introns (10.3%) (Figure S10A) and were also significantly more likely to occur closer to the transcription start site (mean intron position relative to transcript length for introns <100 bp = 24.2% and introns >100 bp = 39.5%; independent-samples t-test t=-54.0, p<0.01) (Figure S10B). Caution should be taken not to overinterpret any differences between short and long introns, as the relationship between intron length and the proportion of CE sites (Figure 9C) is likely driven by shorter introns containing fewer non-constrained sites relative to longer introns (as opposed to shorter introns containing more constrained sites overall). Nonetheless, the enrichment of shorter introns at the start of genes may be worthy of further attention for any possible functional implications on gene regulation.

Finally, we identified 5,611 ultraconserved elements (UCEs) spanning 356.0 kb of the *C. reinhardtii* genome, defined as sequences >=50 bp exhibiting 100% sequence conservation across the three *Chlamydomonas* species. A subset of just 55 UCEs exhibited >=95% sequence conservation across all eight species, indicating that hardly any sequence is expected to be conserved to this level across the core-*Reinhardtinia*. The vast majority of UCE sites (96.0%) overlapped CDS, indicating constraint at both nonsynonymous and synonymous sites. There are several reasons why synonymous sites may be subject to such strong constraint, including interactions with RNA binding proteins, the presence of exonic regulatory elements, or selection for optimal codon usage. Noticeably, 15 of the 55 core-*Reinhardtinia* UCEs overlapped ribosomal protein genes, which are often used as a standard for identifying optimal codons given their extremely high gene expression (Sharp and Li 1987), and several of the other genes overlapped by UCEs are also expected to be very highly expressed (e.g. elongation factors) (Dataset S9). Although considered to be a very weak evolutionary force, this raises the possibility that coordinated selection for optimal codons across the core-*Reinhardtinia* may be a driver of extreme sequence conservation. Alternatively, many of the UCEs may be the result of RNA binding constraints. For example, certain ribosomal proteins may bind and autoregulate their own mRNA (Müller-McNicoll et al. 2019). UCEs have proved to be excellent phylogenetic markers across several taxa (Faircloth et al. 2012; Faircloth et al. 2015). Given the lack of nuclear markers and the current difficulty in determining phylogenetic relationships in the core-*Reinhardtinia*, the 55 deeply conserved elements could potentially be used to provide additional phylogenetic resolution.

3.5 Conclusions

Via the assembly of highly contiguous and well-annotated genomes for three of *C. reinhardtii*'s unicellular relatives, we have presented the first nucleotide-level comparative genomics framework for this important model organism. These resources are expected to enable the continued development of *C. reinhardtii* as a model system for molecular evolution. Furthermore, by providing insights into the gene content and genomic architecture of unicellular core-*Reinhardtinia* species, they are also expected to advance our understanding of the genomic changes that have occurred during the transition to multicellularity in the TGV clade.

Despite such advances, these genome assemblies have only now raised *C. reinhardtii* to a standard that had been achieved for most other model organisms ten or more years ago. Many of the analyses we have performed could be greatly enhanced by the inclusion of additional *Chlamydomonas* species, however addressing this is a question of taxonomy rather than sequencing effort. This is somewhat analogous to the past situation for *Caenorhabditis*, where only very recent advances in ecological knowledge have led to a rapid increase in the number of sampled species and sequenced genomes (Stevens et al. 2019). We hope that this study will encourage the *Chlamydomonas* community to increase sampling efforts for new species, fully enabling the power of comparative genomics analyses to be realised for the species.

3.6 Methods

3.6.1 Nucleic acid extraction and sequencing

Isolates were obtained from the SAG or CCAP culture centres, cultured in Bold's Basal Medium, and where necessary made axenic via serial dilution, plating on agar, and isolation of single algal colonies. High molecular weight DNA was extracted using a customised extension of an existing CTAB/phenol-chloroform protocol (Note S1). One SMRTbell library (sheared to ~20 kb, with 15-50 kb size selection) was prepared per species, and each library was sequenced on a single SMRTcell on the PacBio Sequel platform. PacBio library preparation and sequencing were performed by Edinburgh Genomics.

DNA for Illumina sequencing was extracted using a phenol-chloroform protocol (Ness et al. 2012). Across all species a variety of library preparations, read lengths, insert sizes and sequencing platforms were used (Table S2). RNA was extracted from 4d liquid cultures using Zymo Research TRI Reagent (product ID: R2050) and the Direct-zol RNA Miniprep Plus kit (product ID: R2070) following user instructions. One stranded RNA-seq library was prepared for each species using TruSeq reagents, and sequencing was performed on the Illumina HiSeq X platform (*C. incerta* 150 bp paired-end, *C. schloesseri* and *E. debaryana* 100 bp paired-end). All Illumina sequencing and library preparations were performed by BGI Hong Kong.

3.6.2 *De novo* genome assembly of *Chlamydomonas incerta*

Note that in the following sections on genome assembly (3.6.2, 3.6.3 & 3.6.4) all software parameters, versions and citations are documented in Tables S6 (*C. incerta*), S7 (*C. schloesseri*) and S8 (*E. debaryana*). Mitochondrial assemblies for all three species were assembled and described by Smith and Craig (2020).

The *C. incerta* genome was assembled from 6.31 Gb of PacBio data, with a mean read length of 7.69 kb and an N50 read length of 13.71 kb (Table S1). To estimate genome size and assess the library for possible contaminants, a preliminary genome assembly was produced using miniasm. The taxonomic origin of the resulting contigs was determined by comparison to the NCBI nucleotide collection database (nt) by BLAST+ megablast, and a taxon-annotated GC-coverage plot was produced using Blobtools. This analysis identified a single low-coverage Proteobacteria contaminant, and all reads mapping to contigs identified as bacterial were filtered out. Canu was run using the genome size of the miniasm assembly (130.8 Mb), and as a precaution the Blobtools pipeline was re-run on the resulting assembly, and any further reads identified as bacterial were removed. In total <1% of reads were identified as bacterial over both filtering steps. Canu was then re-run using the final contaminant-filtered dataset. The resulting assembly underwent three rounds of iterative polishing by mapping the PacBio reads using pbalign and performing error correction using the Arrow module of the GenomicConsensus tool.

Further error-correction was performed using ~86x coverage of genomic short-read data and 8.20 Gb of RNA-seq data. The genomic short-read data consisted of two libraries of 100 bp paired-end reads with an insert size of ~180 bp (i.e. overlapping read pairs), and two mate-pair libraries of 100 bp reads with an insert size of ~5,000 bp (Table S2). The short insert size libraries were pre-processed by trimming low-quality bases and adapter sequence using the BBtools program bbduk.sh, mapping the trimmed reads to the Arrow-polished assembly using BWA-MEM, and filtering putative PCR duplicates using Picard MarkDuplicates. For each of the two libraries, the resulting read pairs were then merged using bbmerge-auto.sh to create single reads from the overlapping read pairs where possible. For each library, this analysis resulted in one dataset of unpaired reads (i.e. merged read pairs) and one dataset of paired reads (read pairs that could not be merged). The mate-pair libraries were trimmed of junction adapters and classified as genuine mate-pairs, paired-end, or unknowns using NxTrim. For each library and each classification, pre-processing was performed as described above. Final classification as genuine mate-pairs (i.e. reads pairs with outward facing orientations) or normal paired-end (reads with inward facing orientations) was achieved by assessing the read orientation in the appropriate BAM file, resulting in one dataset of long-insert mate-pairs and one dataset of short-insert paired-end reads per library. All genomic short-read datasets were mapped to the Arrow-polished assembly using BWA-MEM prior to polishing. The RNA-seq dataset consisted of a single library of stranded 150 bp paired-end reads (Table S5). Quality and adapter trimming were performed with Trimmomatic, and reads were mapped to the Arrow-polished assembly using STAR in 2-pass mode. To perform error correction, Pilon was then run by providing all BAM files of aligned short-reads (2x merged single reads and 2x unmerged paired-end reads from the short insert libraries, 2x mate-pair reads and 2x paired-end reads from the long insert libraries, and 1x RNA-seq paired-end reads). The option "--fix bases" was used to avoid genuine introns being removed due to the spliced mapping of the RNA-seq data. Even with this option, we noticed that Pilon corrected a number of large indels, and upon manually checking a sub-sample of such cases using IGV we found that the majority of indels were not supported by the PacBio reads and appeared to be caused by multiply mapping Illumina reads (i.e. the indels appeared to be heterozygous in the Illumina data). We therefore used a custom Perl script to restore all indels >5 bp to their unpolished form after running Pilon. Pilon was run iteratively three times, with all short-read data remapped using BWA-MEM/STAR between each iteration.

The optimal number of polishing iterations for Arrow and Pilon was determined using two metrics: the number of complete single copy orthologs identified by BUSCO (genome mode, Eukaryota odb9 dataset), and the sequence similarity between an existing *C. incerta* EST library (Popescu et al. 2006) and the contigs, as assessed by megablast. Polishing was deemed complete when there was no further increase in either metric between iterations.

Final processing was performed by removing any contigs supported by only a single PacBio read (i.e. "reads=1" in the Canu fasta definition line) and by filtering the two mitochondrion and plastid contigs. Contigs were ordered by size and given unique IDs with the format CXXXX, where XXXX represent ordered numbers (i.e. the largest contig was named C0001). Potential misassemblies were identified via synteny analysis to *C. reinhardtii* (3.6.7). All breakpoints between synteny blocks on a given contig that resulted in a transition between *C. reinhardtii* chromosomes were checked manually using IGV and alignments of the PacBio reads. This resulted in four contigs being split due to likely misassemblies, with split contigs having either "a" or "b" appended to their contig ID.

To identify putative plastid contigs, we used megablast to query the polished Canu assembly with several *C. reinhardtii* plastid gene sequences (*rbcL*, *atpB*, *atpE*, *chlL*, *psbA*, *rrnL* and *ycf4*). A single ~220 kb contig was identified that represented the entire plastid genome, with redundant regions present at the flanks due to assembly as a linear molecule. To produce a circular assembly, we provided Circlator with the putative plastid contig and all Canu error-corrected PacBio reads that mapped to the contig. The resulting circular assembly was then iteratively polished first with Arrow using the raw PacBio reads, followed by Pilon using only the merged paired-end reads. Polishing was performed with each tool until no further changes were introduced, which resulted in two rounds of Arrow and one round of Pilon polishing. The final plastid assembly was a single circular chromosome 184,074 bp in length. The plastid assembly was orientated with *petA* starting at 1 bp.

3.6.3 *De novo* genome assembly of *Chlamydomonas schloesseri*

The *C. schloesseri* genome was assembled from 6.18 Gb of PacBio data, with a mean read length of 7.17 kb and an N50 read length of 12.52 kb (Table S1). Preliminary assembly and contaminant assessment were performed as per *C. incerta* (3.6.2). No putative contaminant contigs were found for the miniasm assembly, although two were identified for the initial Canu assembly. Reads mapping to these contigs were filtered out and Canu was subsequently re-run. Three iterative rounds of polishing were performed with Arrow.

Error correction was performed with Pilon, using ~71x coverage of genomic Illumina data (four libraries, 125 bp paired-end, Table S2) and 7.42 Gb of RNA-seq (one library, stranded 100 bp paired-end, Table S5). Low quality bases and adapter sequences were trimmed using bbduk.sh (genomic data) or Trimmomatic (RNAseq), and putative PCR duplicates were removed from the genomic data using picard MarkDuplicates. The resulting datasets were mapped to the Arrow-polished assembly using BWA-MEM (genomic data) or STAR (RNA-seq), and the resulting BAM files were passed to Pilon. Illumina based polishing was iterated three times.

Final processing was performed by removing any contigs supported by only a single PacBio read and by filtering the mitochondrion and plastid contigs. Potential misassemblies were identified via synteny analysis to *C. reinhardtii* (3.6.7) as per 3.6.2. This resulted in two contigs being split due to likely misassemblies. Contigs were named as per 3.6.2.

A single ~211 kb plastid contig was identified from the polished Canu assembly as per 3.6.2. This contig and Canu error corrected PacBio reads mapping to the contig were passed to Circlator. Polishing was performed per 3.6.2, with no changes observed after the first iteration of Arrow (i.e. no further polishing was performed with Pilon). The final plastid assembly was a single circular chromosome 198,391 bp in length, which was orientated with *petA* starting at 1 bp.

3.6.4 *De novo* genome assembly of *Edaphochlamys debaryana*

The *E. debaryana* genome was assembled from 5.70 Gb of PacBio data, with a mean read length of 7.82 kb and an N50 read length of 13.46 kb (Table S1). Using Blobtools as described above, no contaminant contigs were detected in either the preliminary miniasm assembly or the subsequent Canu assembly. We therefore proceeded with the initial Canu assembly, which was polished with Arrow (three iterations) using all available PacBio reads.

Error correction was performed with Pilon using ~43x coverage of genomic Illumina data (one library, 150 bp paired-end, Table S2) and 7.48 Gb of RNA-seq (one library, stranded 100 bp paired-end, Table S5). Pre-processing was performed as described in 3.6.2 & 3.6.3, and Pilon was run iteratively three times.

Final processing was performed by removing any contigs supported by only a single PacBio read and by filtering the plastid contig (no mitochondrial contigs were detected). Contigs were named as per 3.6.2. We did not check for potential misassemblies given the lack of synteny between *E. debaryana* and *C. reinhardtii* (3.4.4).

A single ~284 kb plastid contig was identified from the polished Canu assembly as described in 3.6.2. This contig and Canu error corrected PacBio reads mapping to the contig were passed to Circlator. Polishing was performed as in 3.6.2, resulting in three rounds of Arrow and one round of Pilon polishing. The final plastid assembly was a single circular chromosome 248,456 bp in length, which was orientated with *petA* starting at 1 bp.

3.6.5 Annotation of genes and repetitive elements

A preliminary repeat library was produced for each species with RepeatModeler v1.0.11 (Smit and Hubley 2008-2015). Repeat models classified as “unknown” with homology to *C. reinhardtii* v5.6 and/or *V. carteri* v2.1 transcripts (e-values <10-3, megablast) were filtered out to remove repeat models that may have been based on gene families. The genomic abundance of each repeat model was estimated by providing RepeatMasker v4.0.9 (Smit et al. 2013-2015) with the filtered RepeatModeler output as a custom library, and any TEs with a cumulative total >100 kb were selected for manual curation, following 1.4.3. Briefly, multiple copies of a given TE were retrieved by querying the appropriate reference genome using megablast, before each copy was extended at both flanks and aligned using MAFFT v7.245 (Katoh and Standley 2013). Alignments were then manually inspected, consensus sequences were created, and TE families were classified following Wicker et al. (2007) and Kapitonov and Jurka (2008). Alongside the most abundant TEs, we also curated *Penelope*-like elements (PLEs) exhaustively using *C. reinhardtii* PLE proteins as queries, which are described in Chapter 5. This procedure was also performed exhaustively for *C. reinhardtii* (i.e. curating all repeat models regardless of genomic abundance), which is also described in Chapter 5. Final repeat libraries were made by combining the RepeatModeler output for a given species with all novel curated TEs (Dataset S1) and *V. carteri* repeats from Repbase. TEs and satellites were softmasked by providing RepeatMasker with the above libraries. In line with the most recent *C. reinhardtii* annotation (Blaby et al. 2014), low-complexity and simple repeats were not masked as the high GC-content of genuine CDS can result in excessive masking.

Adapters and low-quality bases were trimmed from each RNA-seq dataset using Trimmomatic v0.38 (Bolger et al. 2014) with the parameters optimised by Macmanes (2014). Trimmed reads were mapped to repeat-masked assemblies with the 2-pass mode of STAR v2.6.1a (Dobin et al. 2013). Gene annotation was performed with BRAKER v2.1.2 (Hoff et al. 2016; Hoff et al. 2019), an automated pipeline that combines the gene prediction tools Genemark-ET (Lomsadze et al. 2014) and AUGUSTUS (Stanke et al. 2006; Stanke et al. 2008). Read pairs mapping to the forward and reverse strands were extracted using samtools v1.9 (Li et al. 2009) and passed as individual BAM files to BRAKER, which was run with the "—UTR=on" and "—stranded=+,- " flags to perform UTR annotation. Resulting gene models were filtered for genes with internal stop codons, protein sequences <30 amino acids, or CDS overlapped by >=30% TEs/satellites or >=70% low-complexity/simple repeats. Proteins were functionally annotated via upload to the Phycocosm algal genomics portal (Grigoriev et al. 2021).

3.6.6 Phylogenomic analyses

Genome and gene annotations for all available *Reinhardtinia* species and selected outgroups (Datasets S3, S4) were accessed from either Phytozome v12 (if available) or NCBI. For annotation based analyses, protein clustering analysis was performed with OrthoFinder v2.2.7 (Emms and Kelly 2015), using the longest isoform for each gene, the modified blastp options "-seq yes, -soft_masking true, -use_sw_tback" (following Moreno-Hagelsieb and Latimer (2008)) and the default inflation value of 1.5. Protein sequences from orthogroups containing a single gene in all 11 included species (i.e. putative single copy-orthologs) were aligned with MAFFT and trimmed for regions of low-quality alignment using trimAl v1.4.rev15 ("-automated1") (Capella-Gutiérrez et al. 2009). A ML species-tree was produced using concatenated gene alignments with IQ-TREE v1.6.9 (Nguyen et al. 2015), run with ModelFinder ("-m MFP") (Kalyaanamoorthy et al. 2017) and ultrafast bootstrapping ("-bb 1000") (Hoang et al. 2018). ASTRAL-III v5.6.3 (Zhang et al. 2018) was used to produce an alternative species-tree from individual gene-trees, which were themselves produced for each aligned single copy-ortholog using IQ-TREE as described above, with any branches with bootstrap support <10% contracted as recommended.

Annotation-free phylogenies were produced from a dataset of single-copy orthologous genes identified by BUSCO v3.0.2 (Waterhouse et al. 2018) run in genome mode with the pre-release Chlorophyta odb10 dataset (allowing missing data in up to three species). For each BUSCO gene, proteins were aligned and trimmed, and two species-trees were produced as described above.

3.6.7 General comparative genomics and synteny analyses

Basic genome assembly metrics were generated using QUAST v5.0.0 (Gurevich et al. 2013). Repeat content was estimated by performing repeat masking on all genomes as described in 3.6.5 (i.e. supplying RepeatMasker with the RepeatModeler output for a given species plus manually curated repeats from all species). Assembly completeness was assessed by running BUSCO in genome mode with the Eukaryota odb9 and Chlorophyta odb10 datasets. Each species was run with *C. reinhardtii* (-sp chlamy2011) and *V. carteri* (-sp volvox) AUGUSTUS parameters, and the run with the most complete BUSCO genes was retained.

Synteny segments were identified between *C. reinhardtii* and the three novel genomes using SynChro (Drillon et al. 2014) with a block stringency value (delta) of 2. To create the input file for *C. reinhardtii*, we combined the repeat-filtered v5.6 gene annotation (3.6.8, 4.4.8) with the centromere locations for 15 of the 17 chromosomes, as defined by Lin et al. (2018). The resulting synteny blocks were used to check the *C. incerta* and *C. schloesseri* genomes for misassemblies (3.6.2, 3.6.3).

A ML phylogeny of *L1* LINE elements was produced from the endonuclease and reverse transcriptase domains (i.e. ORF2) of all known chlorophyte *L1* elements. Protein sequences were aligned, trimmed and analysed with IQ-TREE as in 3.6.6. All *C. incerta*, *C. schloesseri* and *E. debaryana* elements were manually curated as part of the annotation of repeats (3.6.5). The *Y. unicocca, Eudorina* sp., and *V. carteri* genomes were searched using tblastn with the *ZeppL-1_cRei* protein sequence as query, and the best hits were manually curated as per 3.6.5 to assess the presence or absence of *ZeppL* elements in these species.

3.6.8 Gene annotation metrics and gene family evolution

The *C. reinhardtii* v5.6 gene models were manually filtered based on overlap with the novel repeat library (5.4.1), which resulted in the removal of 1,085 putative TE/repeat genes (4.4.8). For all species, annotation completeness was assessed by protein mode BUSCO analyses using the Eukaryota odb9 and Chlorophyta odb10 datasets. Gene families were identified using OrthoFinder as described in 3.6.6 with the six core-*Reinhardtinia* species with gene annotations (*C. reinhardtii*, *C. incerta*, *C. schloesseri*, *E. debaryana*, *G. pectorale* and *V. carteri*). Protein sequences for all species were annotated with InterPro domain IDs using InterProScan v5.39-77.0 (Jones et al. 2014). Domain IDs were assigned to orthogroups by KinFin v1.0 (Laetsch and Blaxter 2017) if a particular ID was assigned to at least 20% of the genes and present in at least 50% of the species included in the orthogroup.

3.6.9 Mating type locus evolution

Since the *C. reinhardtii* reference genome is *MT*$^+$, we first obtained the *C. reinhardtii MT*$^-$ locus and proteins from NCBI (accession GU814015.1) and created a composite chromosome 6 with an *MT*$^-$ haplotype. A reciprocal best hit approach with blastp was used to identify orthologs, supplemented with tblastn queries to search for genes absent from the annotations. To visualise synteny, we used the MCscan pipeline from the JCVI utility libraries v0.9.14 (Tang et al. 2008), which performs nucleotide alignment with LAST (Kiełbasa et al. 2011) to identify orthologs. We applied a C-score of 0.99, which filters LAST hits to only reciprocal best hits, while otherwise retaining default parameters. We manually confirmed that the LAST reciprocal hits were concordant with our blastp results.

I_{TE} was calculated for each gene using DAMBE7 (Xia 2018). A reference set of highly expressed genes for each species was delineated by performing correspondence analysis on codon usage as implemented in CodonW (http://codonw.sourceforge.net) and taking the default 5% of genes from the extreme of axis 1 (after checking that this set was enriched for genes expected to be highly expressed e.g. histones and ribosomal proteins). The codon usage for the highly expressed reference genes was then provided to DAMBE7, and I_{TE} was calculated for the CDS of each gene using the default option “break 8-fold and 6-fold families into 2”. For both *C. incerta* and *C. schloesseri*, *MID* was annotated by hand as it was absent from the BRAKER annotations (likely due to its short length and unusual codon usage).

3.6.10 Whole-genome alignment and estimating divergence

An 8-species core-*Reinhardtinia* WGA was produced using Cactus (Armstrong et al. 2019) with all available high-quality genomes (*C. reinhardtii* v5, *C. incerta*, *C. schloesseri*, *E. debaryana*, *G. pectorale, Y. unicocca, Eudorina sp.* and *V. carteri* v2). The required guide phylogeny was produced by extracting alignments of 4D sites from single-copy orthologs identified by BUSCO (genome mode, Chlorophyta odb10 dataset). Protein sequences of 1,543 BUSCO genes present in all eight species were aligned with MAFFT and subsequently back-translated to nucleotide sequences. Sites where the aligned codon in all eight species contained a 4D site were then extracted (250,361 sites), and a guide-phylogeny was produced by supplying the 4D site alignment and topology (Figure 2) to phyloFit (PHAST v1.4) (Siepel et al. 2005), which was run with default parameters (i.e. GTR substitution model).

Where available the R domain of the mating type allele not included in a given assembly was appended as an additional contig (extracted from the following NCBI accessions: *C. reinhardtii MT*$^-$ GU814015.1, *G. pectorale MT*$^+$ LC062719.1, *Y. unicocca MT*$^-$ LC314413.1, *Eudorina sp. MT* male LC314415.1, *V. carteri MT* male GU784916.1). All genomes were softmasked for repeats as per 3.6.5, and Cactus was run using the guide-phylogeny and all genomes set as reference quality. Post-processing was performed by extracting a multiple alignment format (MAF) alignment with *C. reinhardtii* as the reference genome from the resulting hierarchical alignment (HAL) file, using the HAL Tools command hal2maf (v2.1) (Hickey et al. 2013), with the options –onlyOrthologs and –noAncestors. Paralogous alignments were reduced to one sequence per species by retaining the sequence with the highest similarity to the consensus of the alignment block, using mafDuplicateFilter (mafTools suite v0.1) (Earl et al. 2014).

Final estimates of putatively neutral divergence were obtained using a method adopted from Green et al. (2014). For each *C. reinhardtii* protein-coding gene, the alignment of each exon was extracted and concatenated. For the subsequent CDS alignments, a site was considered to be 4D if the codon in *C. reinhardtii* included a 4D site, and all seven other species had a triplet of aligned bases that also included a 4D site at the same position (i.e. the aligned triplet was assumed to be a valid codon, based on its alignment to a *C. reinhardtii* codon). The resulting alignment of 1,552,562 sites were then passed to phyloFit with the species tree, as described above.

3.6.11 Identification of false positive and missing genes in *Chlamydomonas reinhardtii*

Genes were first split into control (ortholog, protein domain or both) and test (neither ortholog nor protein domain) datasets. PhyloCSF scores were obtained by passing per exon CDS alignments extracted from the WGA to PhyloCSF (Lin et al. 2011), which was run in “omega” mode using the neutral branch length tree obtained from phyloFit (3.6.10). Following Abascal et al. (2018), the per-gene score was taken as the highest scoring exon, since a small section of misalignment or incorrect annotation (which may be localised to a single exon) can cause an overall negative score for the entire CDS of a genuine protein-coding gene. Exon alignments were trimmed to codon boundaries to preserve reading frame and only exons of at least 45 bp were analysed. If no suitable exons were available for a given gene, the score was taken from the entire CDS. Genetic diversity was calculated from re-sequencing data of 17 *C. reinhardtii* field isolates from Quebec (sampled 1993/94), based on the variant calling and filtering as described by Craig et al. (2019). I_{TE} was calculated for each gene as described above for *C. incerta* and *C. schloesseri*. The Kozak consensus sequence logo for *C. reinhardtii* was determined with WebLogo 3 (Crooks et al. 2004) by providing the 5 bp up- and downstream of the start codons of a randomly selected half of the control gene set (7,682 genes). Kozak scores were calculated for low coding potential genes (i.e. genes that failed all three coding potential tests, 3.4.8), the other half of the control set genes and 10,000 random sequences based on an average genome-wide GC content (64.1%). Following Cross (2015), the score was calculated by summing the per-base bit score from the consensus sequence for each matching base in the query sequence over the 10 sites (i.e. the start codon itself was excluded).

De novo gene annotation was performed on the *C. reinhardtii* v5 genome using BRAKER (without UTR annotation) and all RNA-seq datasets produced by Strenkert et al. (2019). Potential novel genes were defined as those without any overlap with CDS of v5.6 genes. SynChro was re-run against *C. incerta* and *C. schloesseri* using updated *C. reinhardtii* input files containing the potential novel genes. PhyloCSF scores were obtained as above, except scores were taken from entire CDS to ensure only the highest confidence models were retained. BRAKER genes were retained if they had a syntenic ortholog or a PhyloCSF score >100.

3.6.12 Conserved element identification and analyses

CEs were identified from the 8-species WGA using phastCons (Siepel et al. 2005) with the phyloFit neutral model (3.6.10) and the standard UCSC parameters "--expected-length=45, --target-coverage=0.3, --rho=0.31". Parameter tuning was attempted, but it proved difficult to achieve a balance between overly long CEs containing too many non-constrained bases at one extreme, and overly fragmented CEs at the other, and the standard parameters were found to perform as adequately as others.

C. reinhardtii site classes were delineated using the repeat-filtered v5.6 annotation (3.6.8), augmented with the 142 novel genes identified (3.6.11). To assess the genomic distribution of conserved bases, site classes were called uniquely in a hierarchical manner, so that if a site was annotated as more than one site class it was called based on the following hierarchy: CDS, 5' UTR, 3' UTR, intronic, intergenic. Overlaps between site classes and CEs were calculated using BEDtools v2.26.0 (Quinlan and Hall 2010). For analyses of intron length and conservation, all introns were called based on longest isoforms as they appear in the annotation (i.e. no hierarchical calling was performed as described above).

3.7 Acknowledgements

We thank two anonymous reviewers for their insightful comments on an earlier version of the Chapter. We are indebted to Lewis Stevens, Dominik Laetsch and Mark Blaxter for their guidance and advice on all genomics matters. We thank Alexander Suh and Valentina Peona for their invaluable guidance on TE curation, Thomas Pröschold for providing isolate images and for useful discussions on taxonomy, and Olivier Vallon for useful discussions on chlorophyte genomics. We thank Susanne Kraemer and Jack Hearn for their work on an earlier version of the *C. incerta* genome. PacBio sequencing was funded by a NERC Biomolecular Analysis Facility Pilot Project Grant (NBAF1123).

Chapter 4

The Chlamydomonas Genome Project, Version 6: Near Complete Genome Assemblies for Mating Type *Plus* and *Minus* Strains Reveals Extensive Misassembly in Version 5 and Structural Mutation in the Laboratory

4.1 Preface

The work in this chapter was performed as part of a large collaborative effort between researchers from UC Berkeley, the United States Department of Energy, and elsewhere, and the first-person plural is used throughout. All sample preparation, nucleic acids extraction and sequencing were performed at the Joint Genome Institute. Contig-level genome assemblies were produced by Jerry Jenkins at the HudsonAlpha Institute for Biotechnology. I produced chromosomal-level assemblies in collaboration with Olivier Vallon at the CNRS. Preliminary gene annotations were produced by Shengqiang Shu at the Joint Genome Institute. I performed all post-processing of the gene annotations as described in this chapter. One section of the forthcoming manuscript detailing linkage support for the chromosomal assemblies was performed entirely by Patrice Salomé and has been largely omitted. The data underlying Figure 6 were produced by Sean Gallaher. The work was supervised by Sabeeha Merchant and Jeremy Schmutz. I wrote all text, performed all downstream analyses (unless stated above) and produced all figures and tables. General references to work in Chapters 2 and 3 are cited as Craig et al. (2019) and Craig et al. (2021a), respectively.

4.2 Abstract

Over the last two decades, the Chlamydomonas Genome Project has produced five iterations of the *Chlamydomonas reinhardtii* reference genome. The current v5 assembly and annotation were released in 2012, and advances in sequencing technology present a major opportunity to improve these fundamental resources. We produce PacBio-based assemblies for both the mating type *plus* (*MT*$^{+}$) long-term reference strain CC-503 and the mating type *minus* (*MT*$^{-}$) strain CC-4532. Chromosome-level assemblies were produced *de novo* via reference to extrinsic sequencing data, historical linkage data and new knowledge of centromeres and subtelomeres. Assembly contiguity was improved by an order of magnitude and both assemblies are expected to be near complete with respect to genic sequence, with <100 remaining gaps located in the most repetitive genomic regions. Nearly 80% of filled gaps were in genic sequence, providing substantial scope for annotation improvement. The new assemblies revealed extensive intra- and inter-chromosomal misassemblies in v5. Additionally, we found that the CC-503 genome harboured major structural mutations, including a reciprocal translocation, an ~500 kb inversion, and >50 deletions affecting ~100 genes. We therefore recommend the use of CC-4532 as the primary reference, although we also discovered that this strain is experiencing a rapid proliferation of transposable elements (TEs). These results imply that all laboratory strains are expected to harbour several unique structural mutations. Finally, using Iso-Seq and extensive RNA-seq datasets we performed *de novo* structural annotation on each assembly, substantially updating and improving upon past annotations. Collectively, these resources herald an exciting new era of *Chlamydomonas* genomics and are expected to provide the foundation of research in this important model system over the coming years.

4.3 Introduction

As introduced in 1.3.1, five reference assembly versions of the *C. reinhardtii* genome have been produced to date, all of which were primarily based on Sanger sequencing, with additional 454 sequencing performed for v5 (Grossman et al. 2003; Merchant et al. 2007; Blaby et al. 2014). Released in 2012, the v5 assembly spans 111.1 Mb, with ~3.7% gaps, 37 unplaced scaffolds and a contig-level N50 of ~220 kb. Although these assembly metrics represented a considerable achievement at the time of its release, recent developments in long read sequencing technologies have provided the platform to achieve substantially more contiguous assemblies. This has recently been realised for close relatives of *C. reinhardtii,* where Pacific Biosciences (PacBio) sequencing has produced assemblies more contiguous than v5 for five species of unicellular and multicellular volvocine algae (Hamaji et al. 2018; Craig et al. 2021a). Notably, none of these assemblies are chromosome-level, although by combining PacBio sequencing with additional approaches (e.g. optical mapping) highly contiguous chromosome-level assemblies have been produced for more distantly related alga such as *Chromochloris zofingiensis* (Roth et al. 2017). Furthermore, O'Donnell et al. (2020) produced an unannotated assembly of the *C. reinhardtii* laboratory strain CC-1690 (=21 gr) using ultra-long Nanopore sequencing (Liu et al. 2019), in which the 17 chromosomes are assembled as only 21 contigs.

Perhaps of greater significance than assembly contiguity, two recent studies have highlighted inconsistencies between genetic mapping results and the v5 assembly, potentially indicating misassemblies. Salomé and Merchant (2019) presented the case of the phytoene synthase gene *PSY* that is mutated in the white mutant *lts1*, which is currently located on chromosome 2 despite being previously mapped to chromosome 11 (McCarthy et al. 2004). Ozawa et al. (2020) characterised the octotricopeptide repeat protein *MTHI1* that is mutated in the non-photosynthetic mutant *ac46*, and observed that the gene is located on chromosome 17 despite having long since been mapped to chromosome 15 (Dutcher et al. 1991). Interestingly, both inconsistencies were introduced during the transition from v4 to v5, raising the possibility that some of the assembly improvements may have come at the cost of errors.

In addition to the possibility of misassemblies, there is a further potential issue with all previous assembly versions. To meet the high DNA yield requirements of the early genome project, a cell wall-less mutant was used as the reference strain. This strain, CC-503 (=*cw92*), was derived from the *MT*$^+$ strain CC-125 (=137c), which is a “wild type” laboratory strain from the Ebersold/Levine subline (1.2.3). The cell wall-less phenotype was induced by mutagenesis with the methylating agent N-methyl-N’-nitro-N-nitrosoguanidine (MNNG) by Hyams and Davies (1972), although the mutation underlying the phenotype has never been characterised. MNNG primarily induces G:C to A:T transitions, although high doses can induce double strand breaks (DSBs) and chromosomal aberrations (Kaina 2004; Wyatt and Pittman 2006). The fact that CC-503 likely contains at least one non-functional gene has been tolerated based on the expectation that this would have been the result of a point mutation (similar to the unrelated mutations in *NIT1* and *NIT2* that prevent nitrate utilisation and are present in CC-125, and therefore also CC-503 and the reference genome). However, as none of the genetic or molecular mapping results were derived from CC-503, it is possible that some of the inconsistencies between mapping and the assembly may in fact be caused by rearrangements unique to CC-503. Such a hypothesis has never been tested.

Beyond the structure of the assembly itself, the most important source of information are the structural annotations that define the locations of genes, and the protein sequences they encode. As documented in 1.3.1, the annotations have also been regularly updated in line with the reference genome, utilising improvements in annotation algorithms, the availability of homology data to *Volvox carteri* proteins (Prochnik et al. 2010), and most importantly the transition from Sanger- and 454-sequenced expressed sequence tags (ESTs) to deep transcriptomic sequencing (i.e. RNA-seq) performed by Illumina sequencing. Although the current v5 annotation (v5.6) is generally considered to be of high quality, sequencing advances once again present the prospect of further improvements. For example, long read sequencing has been applied to full-length cDNAs, enabling complete isoforms to be captured by single reads and providing unprecedented resolution of gene structures. PacBio cDNA libraries (i.e. Iso-Seq) were recently sequenced for *C. reinhardtii*, leading to the discovery of polycistronic gene expression in the species (Gallaher et al. 2021). Several recent studies have also highlighted specific features of the v5 annotations that could be improved. Via the comparison of a *de novo* transcriptome assembly to the v5 genome, Tulin and Cross (2016) identified more than 100 "hidden" exons within assembly gaps, indicating that increases to contiguity are expected to improve annotation. In a second study, Cross (2015) showed that over 4,000 gene models have in-frame upstream open reading frames (ORFs), and that a substantial proportion of these are likely to represent genuine N-terminal extensions to the proteins encoded by these genes. Blaby and Blaby-Haas (2017) reported that over 100 genes present in annotation v4.3 were not successfully transferred to v5 annotations, including highly expressed and well characterised genes such as *psbW*. Using a comparative genomics approach, Craig et al. (2021a) identified 142 missing genes, recovering 37 of the "lost" v4 genes and finding 25 new genes that were present on polycistronic transcripts together with existing v5.6 genes (Gallaher et al. 2021). As only 87 polycistronic loci were identified in total, this raises the possibility that previous annotations may have overlooked secondary ORFs. Finally, Craig et al. (2021a) performed a thorough quality assessment of v5.6 gene models, reporting that several hundred genes have low coding potential and likely represent false positive models.

Here we present the first major update to both the *C. reinhardtii* assembly and annotation in over eight years. We first present a PacBio-based *de novo* assembly for CC-503, which we use to document the correction of several large misassemblies and to describe the sequence context of assembly gaps in v5 that have now been filled. However, our updated assembly revealed that the CC-503 genome harbours many large structural mutations predicted to affect ~100 genes. We also present a PacBio-based *de novo* assembly of the *MT*$^{-}$ laboratory strain CC-4532, and recommend that this assembly be used as the primary reference genome in most user cases. We present high quality structural annotations for both assemblies, with the incorporation of Iso-Seq data leading to considerable improvements. These updates mark the start of an exciting new era of *Chlamydomonas* genomics, with developing opportunities to produce reference quality assemblies and annotations for several strains and field isolates of the species.

4.4 Results and Discussion

4.4.1 CC-503 version 6: a long read *Chlamydomonas reinhardtii* reference genome

As the first stage in updating the reference genome, we produced a *de novo* contig-level assembly for CC-503 from a high coverage (~127x) long read dataset sequenced on the PacBio Sequel platform (4.5.1). In line with the reported inconsistencies (4.3), we identified several contradictions between the v5 chromosomes and the *de novo* assembled contigs. Consequently, we entirely reassembled all well-supported contigs to chromosomes, without any reference to previous versions. This was achieved by integrating evidence from several data sources. Primarily, we mapped the contigs to the highly contiguous Nanopore-based CC-1690 assembly (O'Donnell et al. 2020). The resultant map of *de novo* assembled contigs to CC-1690 was then further validated against the PacBio *de novo* assembled contigs of CC-4532 (4.4.5). This approach not only enabled contigs to be placed on chromosomes in a manner consistent across all three assemblies (with the exception of chromosomes 2 and 9, 4.4.4), but also for gap lengths between remaining contig breaks to be estimated relative to CC-1690. All major structural changes were validated against genetic and molecular mapping data. Finally, recent knowledge of the repetitive structure of centromeric and subtelomeric regions provided extrinsic evidence supporting the validity of the new chromosome models (i.e. that each chromosome contained one centromere and terminated in subtelomeres).

The resulting chromosome-level assembly (CC-503 v6) exhibited dramatic improvements in contiguity relative to previous versions (Table 1). Comparing v5 and CC-503 v6, the number of gaps was reduced by an order of magnitude, decreasing from 1,495 to 145. Consequently, the contig-level N50 increased from 0.22 Mb to 2.92 Mb, with ~85% of CC-503 v6 assembled on contigs >1 Mb. The proportion of unknown bases (i.e. Ns) dropped more modestly from 3.7% to 1.7%, although this figure is inflated in CC-503 v6 by the inclusion of very long estimated gaps (maximum 300 kb) in the most repetitive regions of the genome. Unlike in previous versions, the unplaced sequences were assembled as contigs (as opposed to scaffolds), named contig_18 to contig_60. The cumulative length of unplaced sequence was reduced from 2.2 Mb to 1.5 Mb, and although the number of unplaced sequences is similar between versions (37 scaffolds in v5 and 42 contigs in CC-503 v6), the sequences themselves are largely unrelated. Sequence at least partially corresponding to 33 of the 37 v5 unplaced scaffolds was assembled on chromosomes in CC-503 v6, with the unplaced contigs consisting of the remaining v5 unplaced scaffolds, novel sequence from extremely repetitive regions and a number of regions misassembled on chromosomes in v5 that could not be reliably placed on chromosomes in CC-503 v6. A thorough assessment of the assembly improvements and the remaining problematic regions is presented in the following text.

Table 1. Comparison of assembly metrics between reference genome versions.

Assembly version	CC-503 v 4	CC-503 v5	CC-503 v6	CC-4532 v6
Year of release	2008	2012	/	/
Placed length (Mb)	102.62	108.90	110.10	112.32
Unplaced scaffolds/contigs	71	37	42	40
Unplaced length (Mb)	9.68	2.20	1.45	1.72
Number of contigs	2,739	1,495	145	120
Contig N50 (Mb)	0.09	0.22	2.92	2.65
GC (%)	64.06	64.08	64.07	64.11
Ns (%)	7.54	3.65	1.66	0.92
Transposable elements (%)	9.76	10.52	10.71	12.33
Microsatellite (monomers <10 bp) (%)	1.32	1.44	1.73	1.77
Satellite DNA (monomers >=10 bp) (%)	2.86	3.14	3.87	4.12

4.4.2 The version 6 assembly reveals misassemblies in version 5

CC-503 v6 exhibited major structural differences to v5, which affected the ordering and orientation of sequence both within and between chromosomes. Overall, only six chromosomes (1, 4, 6, 7, 13 and 14) remained entirely consistent with respect to the ordering of scaffolds in v5. The extent of the changes to the remaining 11 chromosomes ranged from the minor intra-chromosomal reordering of short contigs to major inter-chromosomal rearrangements affecting megabases of sequence. An overview of the between chromosome changes is presented in Figure 1A.

Many of the changes occurred in close proximity to the most repetitive genomic regions, notably the putative centromeres, subtelomeres and regions corresponding to unplaced scaffolds in previous versions. As in many species, precise centromeric locations in *C. reinhardtii* have remained elusive. Lin et al. (2018) reported the presence of 200-800 kb regions containing multiple genes encoding proteins with reverse transcriptase domains that flanked molecular markers known to be tightly linked to centromeres. Craig et al. (2021a) showed that these genes are encoded by multiple copies of an *L1* LINE retrotransposon that is homologous to *Zepp*, the putative centromeric component of the distantly related alga *Coccomyxa subellipsoidea* (Blanc et al. 2012). These *Zepp*-like (*ZeppL*) elements are highly localised at the putative centromeres, although chromosomes 2, 3, 5 and 8 contained two clusters per chromosome in v5, and chromosomes 11 and 15 entirely lacked clusters (Lin et al. 2018; Craig et al. 2021a). The *C. reinhardtii* telomeric repeat (TTTTAGGG)$_n$ has been characterised previously (Petracek et al. 1990), while long complex satellite arrays that are highly characteristic of subtelomeric regions have recently been identified and described (Chaux-Jukic et al. 2021).

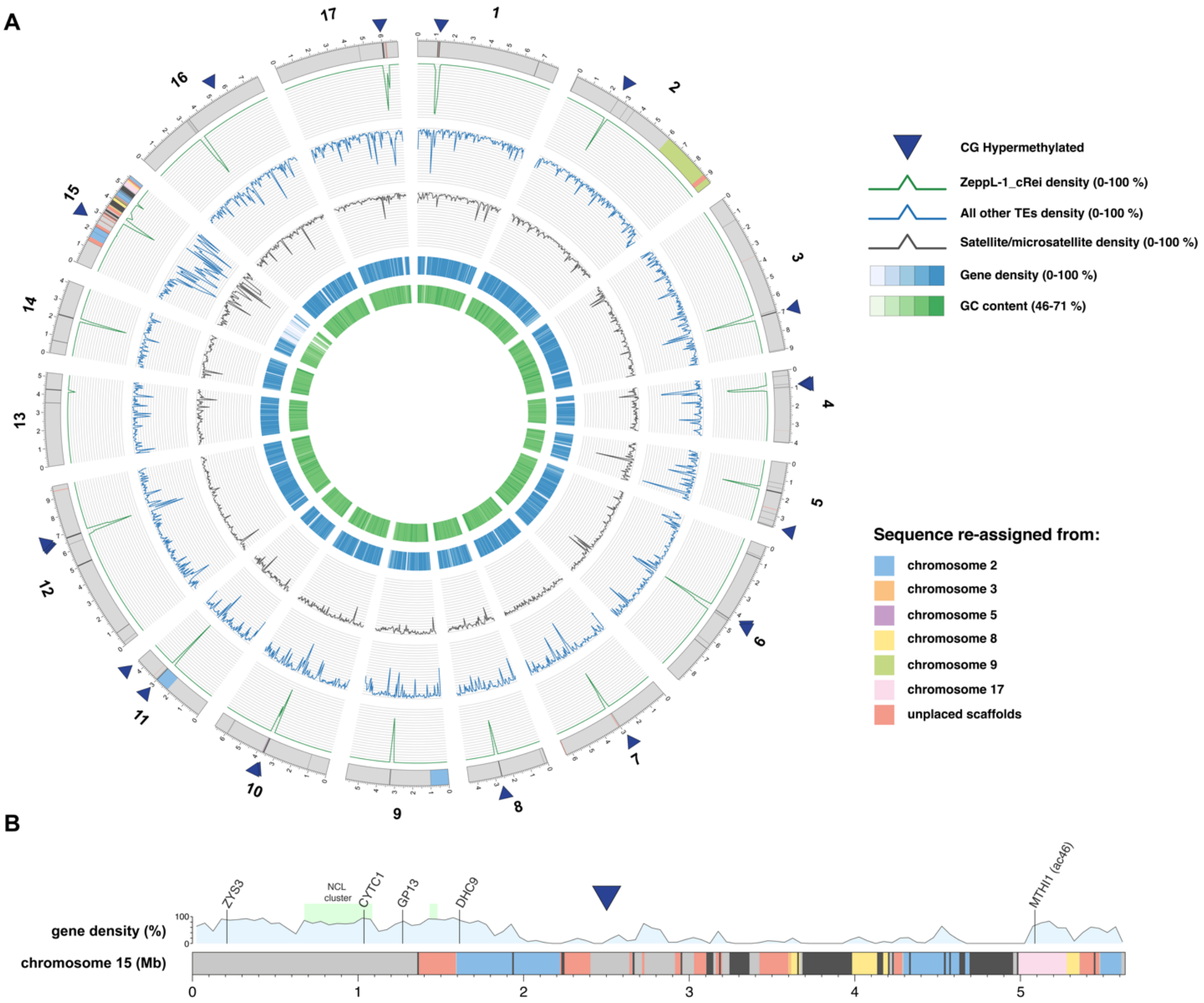


Figure 1. Overview of the CC-503 v6 assembly.
(A) Circos plot (Krzywinski et al. 2009) representation of CC-503 v6. Grey outer blocks represent chromosomes, with additional colours highlighting genomic regions that were assembled on other chromosomes or on unplaced scaffolds in v5. Dark grey regions represent gaps between contigs, with any gaps <10 kb increased to 10 kb to aid visualisation. All metrics were calculated for 50 kb non-overlapping windows. CG hypermethylated regions were taken from Lopez et al. (2015) and mapped from v5 to CC-503 v6.
(B) Linear representation of chromosome 15. Marker genes are from Kathir et al. (2003) and the light green boxes represent the *NCL* gene clusters identified by Boulouis et al. (2015).

Comparisons of the v5 and CC-503 v6 assemblies of chromosomes 5 (Figure 2A) and 11 (Figure 2B) illustrate the types of misassemblies that have affected these regions. In v5, the left arm of chromosome 5 terminates in a 47 kb contig containing *ZeppL* elements (purple block, Figure 2A), which in CC-503 v6 has been assembled within the putative centromere of chromosome 10 (Figure S1D). The remaining regions of chromosome 5, the three fragments of ~0.7, 1.2 and 1.7 Mb (light blue, yellow and orange, respectively), have then been rearranged. The mis-join between the light blue and yellow fragments involved an assembly gap corresponding to scaffold 24 in v5 (containing *MUT6* in Figure 2A), which in CC-503 v6 is assembled between the orange and light blue fragments. The mis-join between the yellow and orange fragments contained a subtelomeric repeat in CC-503 v6, which is now correctly placed after the yellow fragment was reversed. Thus, the reassembled chromosome 5 has subtelomeric repeats at both termini, a single internal *ZeppL* cluster, and is also entirely consistent with the molecular map (Kathir et al. 2003). The movement of an ~770 kb region from chromosome 2 to 11 simultaneously resolved the absence of a putative centromere on chromosome 11 and the presence of two *ZeppL* clusters on chromosome 2 (Figure 2B). This region also includes *PSY*, which as described in 4.3 has been genetically mapped to chromosome 11 (McCarthy et al. 2004; Salomé and Merchant 2019). Independently, an ~860 kb fragment (light blue) was inverted, which is consistent with the chromosomal positions of *PETC* and *VFL2* being indistinguishable from one another in the molecular map (Kathir et al. 2003). Examples of misassemblies affecting other chromosomes are presented in Figure S1.

The most dramatic changes affected chromosome 15, which at 1.92 Mb was the shortest chromosome in v5. In CC-503 v6, chromosome 15 has almost tripled in assembled length to 5.63 Mb, acquiring sequence previously assembled on four chromosomes (2, 3, 8 and 17) and 16 unplaced scaffolds (Figure 1B). For chromosome 2, all sequence from 8.0 Mb onwards (~1.2 Mb total) in v5 was reassigned to three different regions of chromosome 15 (Figure 1B, Figure S1A). One of these regions contained *DHC9*, which had previously been genetically mapped to chromosome 15 (Porter et al. 1996; Kathir et al. 2003). The single 0.29 Mb region reassigned from chromosome 17 contained *MTHI1*, which as described in 4.3 was also genetically mapped to chromosome 15 (Dutcher et al. 1991; Ozawa et al. 2020). Included amongst the regions reassigned from chromosome 8 (0.22 Mb total) and several of the previously unplaced scaffolds (0.90 Mb total) were regions containing *ZeppL* elements, explaining the previous absence of centromeric sequence on chromosome 15. Although only 20 kb was reassigned from chromosome 3 to 15, additional sequence previously assembled on chromosome 3 now corresponds to contig_19 (89 kb) in CC-503 v6 (Figure S1B). Similarly, sequences assembled on chromosomes 2 and 8 in v5 correspond to regions of contig_18 (206 kb) and contig_22 (58 kb) in CC-503 v6 (Figure S1A, C).

The newly assembled chromosome 15 has several features that distinguish it from other chromosomes. It is by far the most repeat-rich and gene-poor, with an average repeat content of 45.3% (mean 15.1% for the other 16 chromosomes) and gene density of 43.4% (mean 80.2% for the 16 other chromosomes) (Table S1). Furthermore, this pattern is not uniform, with the first ~1.9 Mb of the left arm appearing to be relatively normal (repeat content 21.7%, gene density 72.8%) and the remainder of the chromosome being massively repetitive (61.3%) and gene poor (23.2%) (Figure 1B). As a result, chromosome 15 is by far the most fragmented in CC-503 v6, with its 21 contigs and mean estimated gap length of 43.3 kb considerably higher and longer relative to the remaining chromosomes (5.1 contigs, 15.0 kb mean gap length). Given the long gaps and the extreme repeat content of the unplaced contigs (77.5% repeats), we expect that a substantial proportion of the unplaced sequence in CC-503 v6 belongs on chromosome 15. Unfortunately, we were unable to unambiguously place these contigs by mapping between the CC-1690 and CC-4532 v6 assemblies, or by any other source of evidence.

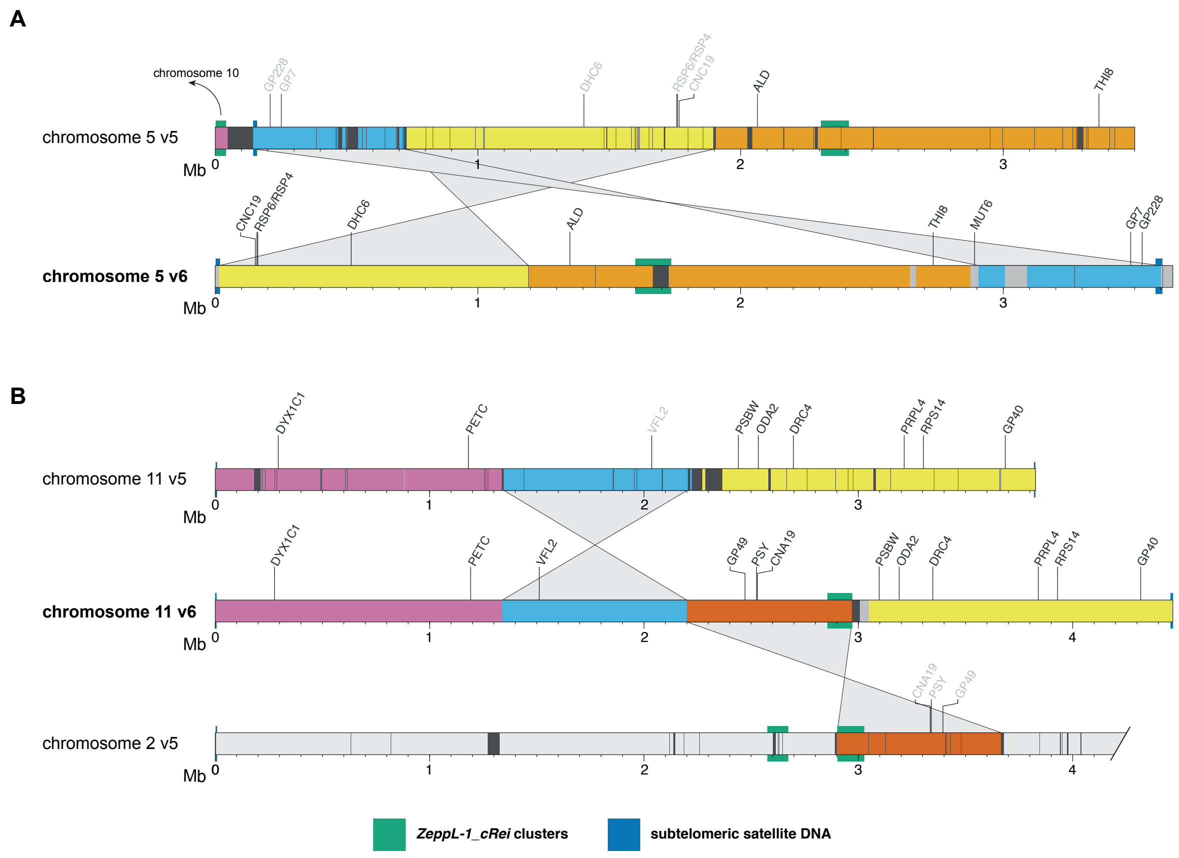


Figure 2. Examples of v5 misassemblies and their resolution in CC-503 v6.
Chromosome segments are coloured to show the reordering and reorientation of specific regions, and dark grey regions represent assembly gaps. Markers inconsistent with the molecular map of Kathir et al. (2003) are shown in light grey text.
(A) Reassembly of chromosome 5, the purple segment has been reassigned to chromosome 10 (Figure S1D). Light grey regions on the CC-503 v6 chromosome correspond to sequence not represented on the v5 chromosome (e.g. the block containing *MUT6* corresponds to scaffold 24 in v5). *RSP4* and *RSP6* are neighbouring genes that correspond to the *pf1* and *pf26* genetic markers, respectively (Dutcher 2014).
(B) Reassembly of chromosome 11, only the first 4.2 Mb of chromosome 2 is shown. Genes that originally corresponded to genetic markers in Kathir et al. (2003) are: *PSY – lts1* (McCarthy et al. 2004), *DYX1C1 – pf23* (Yamamoto et al. 2017b), *DRC4 – pf2* (Dutcher 2014), *PRPL4 – ery1, RPS14 – cry1*.

The unusual features of chromosome 15 raise several questions about its evolutionary origins, gene content and chromosomal environment. Except for *MTHI1*, all chromosome 15 marker genes (*ZYS3*, *CYTC1* and *DHC9*) are located within the relatively gene-rich first ~1.9 Mb of the chromosome. This region is also notable for containing almost all of the *NCL* genes, a gene family of RNA binding proteins that is experiencing an ongoing diversification in *C. reinhardtii* (Boulouis et al. 2015). Of the 38 identified *NCL* genes, 32 are present in a single cluster spanning ~410 kb, while a further three are present in a shorter upstream cluster that was assembled on scaffold 19 in v5 (Figure 1B). The yellow-in-the-dark mutation *y1* has also been mapped to the left arm of chromosome 15 and is linked to *DHC9* (Porter et al. 1996). It is possible that the unknown *Y1* gene was assembled on either chromosome 2 or an unplaced scaffold in v5, and it may be worth revisiting attempts to identify it. The remainder of chromosome 15 contains only 152 annotated genes, although several of these are expected to be essential, for example the plastid 50s ribosomal protein L3 (*PRPL3*), the 26s proteasome regulatory subunit (*RPN9*), and indeed *MTHI1*. Given its repeat content, it would be interesting to determine if much of chromosome 15 is packaged as heterochromatin, and if that is the case, whether genes are expressed from heterochromatic environments (e.g. as is the case for many genes on the repeat-rich dot chromosome in *Drosophila melanogaster* (Riddle and Elgin 2018)). Similarly, it would be interesting to know if the high repeat content results in an atypical recombination landscape on chromosome 15. Unfortunately, little is currently known about genome-wide variation in either heterochromatin or fine-scale recombination rate.

As outlined above for chromosomes 5, 11 and 15, historic linkage data supported the inferred misassembly corrections. Two independent linkage maps were produced for all chromosomes using either all molecular markers from Kathir et al. (2003) or the whole-genome re-sequencing data of tetrad progeny from Liu et al. (2018) (analyses performed by Patrice Salomé, data not shown). Briefly, in a comparison between v5 and CC-1690 (on which CC-503 v6 is based), the total map distance was reduced from 6,676 cM to 1,473 cM for the marker-based map, and from 6,965 cM to 1,363 cM for the re-sequencing-based map. All chromosomes received strong validation from this analysis except for chromosomes 2 and 9 in CC-503 v6. As documented in 4.4.4, this discrepancy was the result of a putative reciprocal translocation between these chromosomes that is unique to CC-503.

4.4.3 Sequence context of filled gaps and the final assembly challenges

Given the substantial increase in contiguity between v5 and CC-503 v6, we aimed to characterise the assembly gaps that have been filled and the genomic regions that have continued to resist assembly. We identified TEs using the updated *C. reinhardtii* TE library (5.4.1) and RepeatMasker (Smit et al. 2013-2015), and microsatellites (tandem repeats with monomers <10 bp) and satellite DNA (monomers ≥10 bp) using Tandem Repeats Finder (Benson 1999) (4.5.2). As a result of filled gaps and newly assembled sequence, CC-503 v6 contains ~2.6 Mb of novel sequence. The proportion of the genome identified as repetitive increased from 15.1% to 16.3% between v5 and CC-503 v6 (Table 1), representing an increase of ~1.4 Mb. However, there was only a modest increase in the proportion of the genome identified as TEs (10.5% vs 10.7%). A more substantial increase was observed between v4 and v5 (9.8% vs 10.5%), indicating that the targeted gap filling performed for v5 was largely successful in assembling TEs. Conversely, there were more substantial increases in the assembly of microsatellites and satellite DNA, with both classes increasing by ~20% relative to v5. Although these sequences represent a relatively small proportion of the genome (1.7% microsatellites and 3.9% satellite DNA in CC-503 v6, Table 1) they appear to have been responsible for a large proportion of the v5 assembly gaps.

Of the 706 gaps that could be unambiguously mapped forward and verified as filled in CC-503 v6, 21.2% contained predominantly microsatellites, 15.6% satellite DNA and 10.1% TEs (Figure 3A). Relative to the CC-503 v6.1 annotation (4.4.7), 22.0% of filled gaps were intergenic, 41.6% were intronic and 36.4% had at least partial overlap with novel exonic sequence. The intronic gaps were particularly enriched for microsatellite and satellite sequences, while the majority of filled intergenic gaps that could be classified to a single repeat class were associated with TEs (Figure 3A). In line with this, only 4.8% of intronic sequence was derived from TEs, compared to 38.37% for intergenic sequence (Table S2). Conversely, 3.5% and 4.7% of intronic sequence was annotated as microsatellite and satellite sequences, respectively, relative to 1.4% and 5.6% for intergenic sequence. To express this with respect to the repetitive sequences themselves, 13.8% of total TE sequence, 61.7% of microsatellite sequence and 37.2% of satellite sequence was found in introns, relative to 78.0%, 17.7% and 31.2% in intergenic sequence. Thus, although introns constitute more than 30% of the genome and nearly 60% of non-exonic sequence, they appear to exhibit an underrepresentation of TE sequence, but an overrepresentation of tandem repeats. These repeats frequently caused assembly gaps in previous versions, and the improvements in contiguity have therefore provided substantial potential to improve gene model annotation (4.4.7).

The CC-503 v6 chromosomes contain only 86 remaining gaps, which are expected to occur at the most repetitive regions of the genome. Approximately half of the gaps are located in putative centromeres (34 gaps) and subtelomeres (seven gaps). Of the remaining gaps, 11 were flanked on both sides by satellite DNA, seven by microsatellites, four by fragments of the same TE, and 23 by different classes of sequence suggesting a more complex structure (Figure 3B, Dataset S1). Considering putative centromeres, each chromosome now contains a single localised cluster of *ZeppL* elements (Figure 1A). The only possible exception was chromosome 15, where the *ZeppL* elements were found more disparately across an ~1.5 Mb region. We estimated the assembly completeness of these regions relative to CC-1690, in which the putative centromeres of all chromosomes (except chromosome 15) are assembled without gaps (O'Donnell et al. 2020). Estimating the length as the span of the *ZeppL* clusters, the putative centromeres range from 75 – 439 kb in CC-1690, with a mean of 205 kb (Figure 3C, Table S3). The proportion of these regions assembled in CC-503 v6 ranged from 36.8% to 99.0%, with 75.1% assembled across all chromosomes. This corresponds to ~870 kb of sequence that is missing from CC-503 v6 chromosomes, and likely the entire assembly since only two unplaced contigs contain *ZeppL* elements. As reported by Craig et al. (2021a), the putative centromeres are also enriched for other TEs (Figure 1A, Figure S2 for CC-1690), with the *ZeppL-1_cRei* element itself comprising ~60% of the putative centromeres (Table S3). Overall, the *ZeppL* clusters accounted for 25.4% of all TE sequence despite spanning only 3.1% of the genome. The exact repeat structure of these regions and the variation that exists between chromosomes warrants further study. Another important step to confirm that these regions are the centromeres will be to determine the localisation of the centromeric histone H3 (CenH3, also CENP-A in many non-plant species), as was recently performed for *D. melanogaster* (Chang et al. 2019).

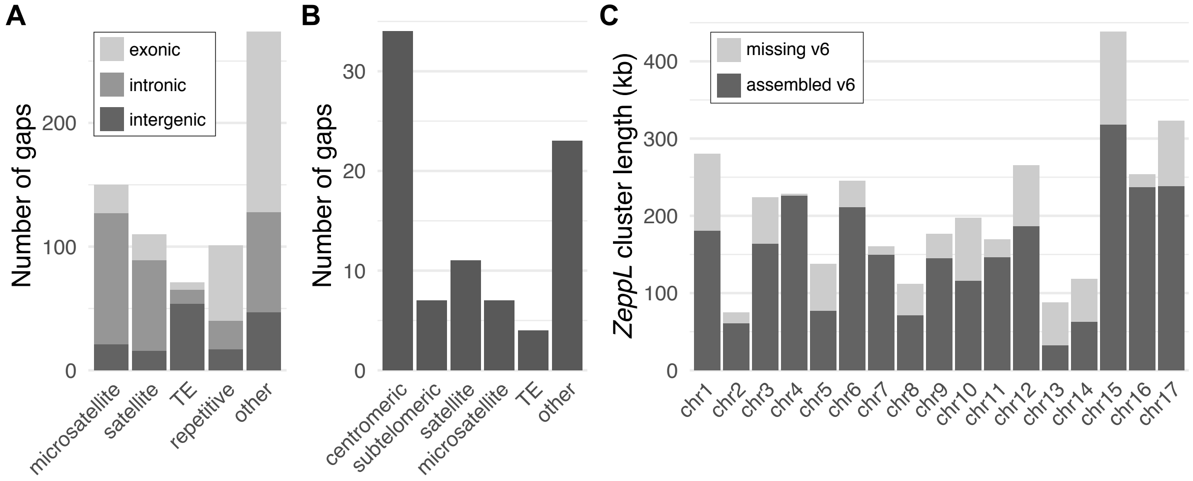


Figure 3. Filled gaps and the remaining assembly challenges in CC-503 v6.
(A) Repeat class of v5 gaps filled in CC-503 v6, bars are split by entirely intergenic gaps, entirely intronic gaps and gaps with at least partial exonic overlap. "Repetitive" refers to gaps with >25% repeat content but no overall majority of repeat type. "Other" filled gaps had repeat contents <25% (4.5.3).
(B) Classification of the remaining gaps in CC-503 v6, gaps classed as "other" contained more than one repeat class at each flank and includes complex regions such as tandem duplications (e.g. the "16 kb repeats", 4.4.5).
(C) Summary of the putatively centromeric *ZeppL* cluster lengths in the CC-1690 assembly, and the proportions of the clusters that are assembled in CC-503 v6.

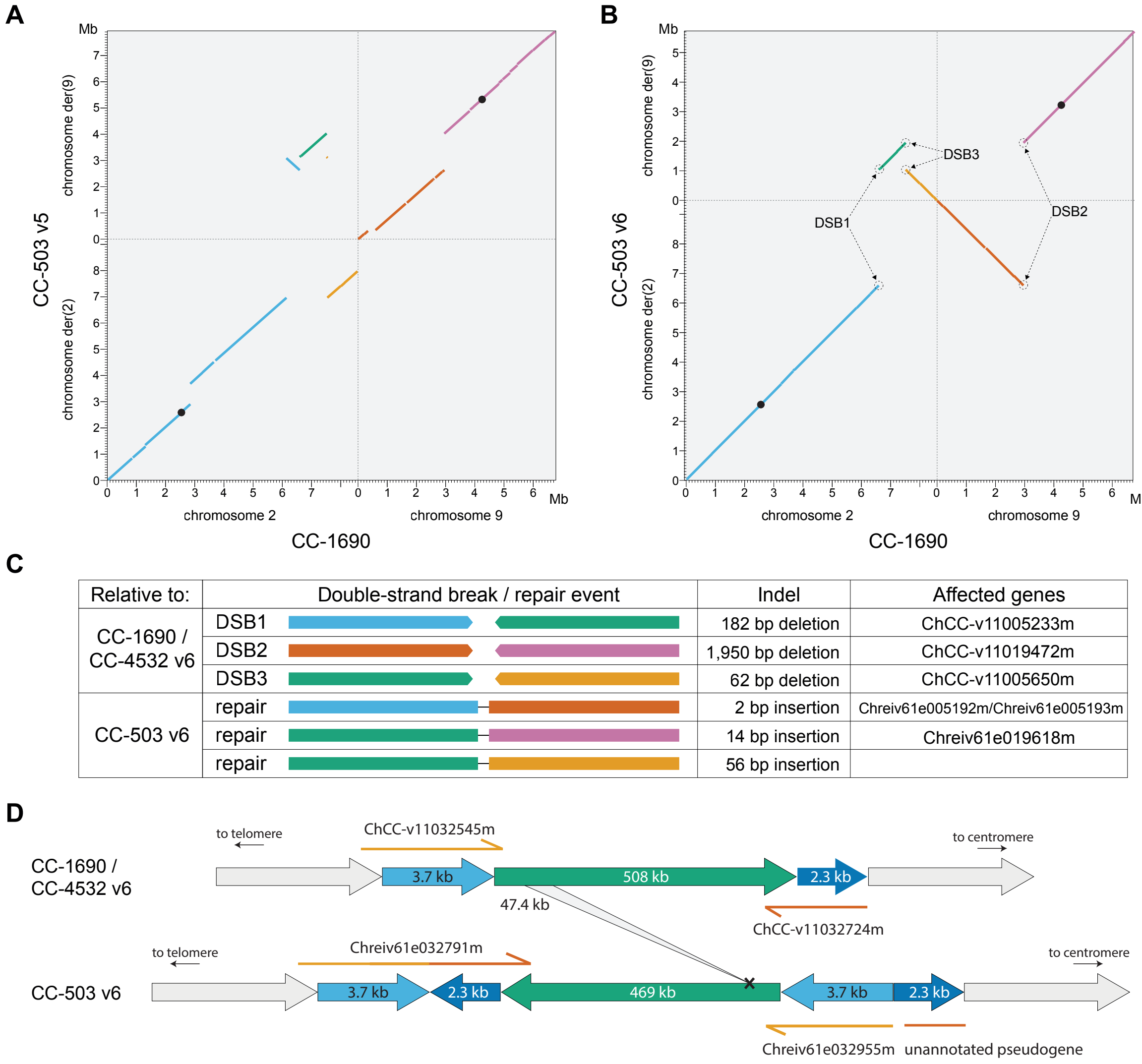
The highly repetitive nature of subtelomeric sequence, especially the presence of complex satellite arrays (Chaux-Jukic et al. 2021), resulted in the truncation of 24 of the 34 chromosome termini (i.e. only 10 chromosome arms terminated in assembled telomeric repeats). The nature of the satellite repeats makes assembly challenging even with long reads, and these regions may need to be targeted with ultra-long reads (as for CC-1690) or alternative technologies. All three long read assemblies (CC-503 v6, CC-1690 & CC-4532 v6) revealed that the ribosomal DNA (rDNA) present at the left arm terminus of chromosome 1 is not a complete array as found at the right arm termini of chromosomes 8 and 14. Instead, the chromosome 1 terminus contains only one disrupted complete rDNA copy and one partial copy immediately upstream of the $(TTTTAGGG)_n$ telomeric repeat (Figure S3).

Finally, we re-assessed the genomic locations of the CG hypermethylated regions identified by Lopez et al. (2015). These regions generally coincided with the putative centromeres, although two were located at subtelomeres on chromosomes 5 and 11, and two were located at non-centromeric/subtelomeric regions on chromosomes 15 and 16 (Figure 1A). Hypermethylation of subtelomeres has been confirmed by Chaux-Jukic et al. (2021), a result that was likely largely missed by the previous hypermethylated regions due to the low assembly quality of these regions in v5 and difficulties in mapping short read bisulfite sequencing data to satellite DNA. Collectively, these results suggest that although CG methylation is generally very low (<1%) at the genome-wide scale (Lopez et al. 2015), hypermethylation occurs in the most densely repetitive regions of the genome.

4.4.4 The CC-503 genome harbours major structural mutations

The *de novo* chromosomal assembly and subsequent linkage analysis revealed a major inconsistency between chromosomes 2 and 9 of CC-503 v6 and CC-1690. Given that the linkage analysis was performed using data from field isolates, and that CC-4532 v6 is entirely consistent with CC-1690 (4.4.5), this suggests that the CC-503 genome contains a genuine structural mutation. To distinguish between the ancestral and mutant states in the following text, the derivative CC-503 chromosomes will be referred to as der(2) and der(9). In the v5 assembly, the aberration is assembled as a 1.4 Mb translocation from chromosome 2 to 9, with the translocated sequence broken into three rearranged fragments relative to CC-1690 (Figure 4A). This rearrangement requires a complex chain of structural mutations and at least five DSBs. Via manual inspection of the CC-503 v6 contigs we were able to infer chromosome assemblies that could be explained by a more parsimonious, although still complex, order of mutational events. Under this model, chromosomes 2 and 9 have experienced a reciprocal translocation, with an additional inversion affecting part of the fragment translocated from chromosome 2 to 9 (Figure 4B). This scenario requires three DSBs, one between the blue and green fragments of chromosome 2 (DSB1), a second between the purple and vermillion fragments of chromosome 9 (DSB2), and a third between the green and orange fragments of chromosome 2 (DSB3). The proposed mis-repair of these DSBs in CC-503 has produced chromosomes der(2) and der(9), with 3.0 Mb from the left arm of chromosome 9 (vermillion fragment) translocated to chromosome 2, and reciprocally 2.0 Mb from the right arm of chromosome 2 (green and orange fragments) translocated to chromosome 9. The 0.9 Mb inversion (green fragment) presumably shares DSB1 with the translocation event, implying that the three DSBs and subsequent rearrangements occurred simultaneously.

All of the DSBs and repair events were associated with indels. Relative to CC-1690, DSB1 resulted in a 179 bp deletion, DSB2 a 1,950 bp deletion and DSB3 a 62 bp deletion (Figure 4C). Re-sequencing data from CC-125 (the progenitor of CC-503) mapped across each DSB and deletion, implying that the rearrangement is unique to CC-503 (Figure S4). The putative repair sites in CC-503 v6 contained short insertions. The repair between the blue and vermillion fragments resulted in a dinucleotide “TA” insertion, the repair between the purple and green fragments a 14 bp insertion, and the repair between the green and orange fragments a 56 bp insertion (Figure 4C). Of the latter insertion, 53 of 56 bp matched perfectly and uniquely to a region within the 1,950 bp deletion at DSB2, supporting the hypothesis that the DSBs and rearrangements occurred simultaneously. Since CC-4532 v6 (4.4.5) is entirely consistent with the unmutated CC-1690 assembly, the CC-4532 v6.1 annotation (4.4.7) can be used to determine if any genes have been disrupted by the rearrangement. Each of the DSBs was predicted to disrupt coding sequence, with the three affected genes listed in Figure 4C and shown by browser views in Figure S5. As an example, DSB1 has entirely deleted an exon of a gene (ChCC-v11005233m) encoding a 318 amino acid protein with an S-adenosylmethioine-dependent methyltransferase domain, with the remaining (and presumably pseudogenised) exons now split between der(2) and der(9) in CC-503 v6 (Figure S5A).



Relative to:	Double-strand break / repair event		Indel	Affected genes
CC-1690 / CC-4532 v6	DSB1	🟦 🟩	182 bp deletion	ChCC-v11005233m
	DSB2	🟧 🟪	1,950 bp deletion	ChCC-v11019472m
	DSB3	🟩 🟨	62 bp deletion	ChCC-v11005650m
CC-503 v6	repair	🟦 🟧	2 bp insertion	Chreiv61e005192m/Chreiv61e005193m
	repair	🟩 🟪	14 bp insertion	Chreiv61e019618m
	repair	🟩 🟨	56 bp insertion	

Figure 4. Structural mutations in the CC-503 genome.
(A) Dotplot representation of chromosomes 2 and 9 between v5 and CC-1690. Colours link fragments between panels **(A)** and **(B)**. Black circles represent putative centromeres.
(B) Dotplot representation of chromosomes 2 and 9 between CC-503 v6 and CC-1690.
(C) Summary of indels present at DSBs and repair points. For DSBs, CC-4532 v6.1 gene IDs are listed to represent the predicted non-mutant genes. For repair points, CC-503 v6.1 gene IDs are listed to represent predicted genes that are expected to be incomplete and likely non-functional as a result of the mutation.
(D) Schematic representation of the dupINVdup/deletion double mutation. The duplicated flanks (light and dark blue) are shown 50x the scale of the main inverted fragment (green). The 47.4 kb internal deletion is represented by the grey ribbon, see Dataset S2 for affected genes. Non-mutant genes are represented by the CC-4532 v6.1 gene IDs, while mutant gene annotations (including a predicted fusion gene on the left flank) are provided by CC-503 v6.1 gene IDs.

Based on this discovery, we performed a thorough assessment of structural variation between CC-503 v6 and CC-4532 v6 in order to detect any smaller aberrations, polarising mutations with CC-1690 (i.e., a mutation was considered to be derived in a given assembly if the other assembly was consistent with CC-1690). Remarkably, we identified 74 structural mutations (excluding TE variants, 4.4.6) unique to CC-503 v6 that are predicted to affect 103 genes

(Dataset S2), including deletions >10 kb that entirely remove several neighbouring genes. In full, we identified 66 deletions (cumulatively 309.5 kb), six insertions (cumulatively 29.0 kb), one inversion and one duplication. In comparison, we identified only seven structural mutations unique to CC-4532 v6, predicted to affect seven genes (Dataset S3). Aside from the reciprocal translocation, the most striking of the CC-503 v6 mutations was an ~508 kb inversion occurring between 0.81 and 1.32 Mb on chromosome 16 (Figure 4D). Inspection of the two DSBs and their subsequent repair revealed that this event is an unusual dupINVdup (duplication-inversion-duplication) mutation (Brand et al. 2015), in which both flanks of the unique inverted fragment are duplicated and themselves inverted. This has resulted in duplications of 3,708 bp and 2,309 bp from the left and right flanks of the uniquely inverted sequence, respectively, both of which have disrupted and partially duplicated genic sequence. Remarkably, the inverted region itself harbours a 47.4 kb deletion which has partially or fully deleted ten genes (Figure 4D, Dataset S2).

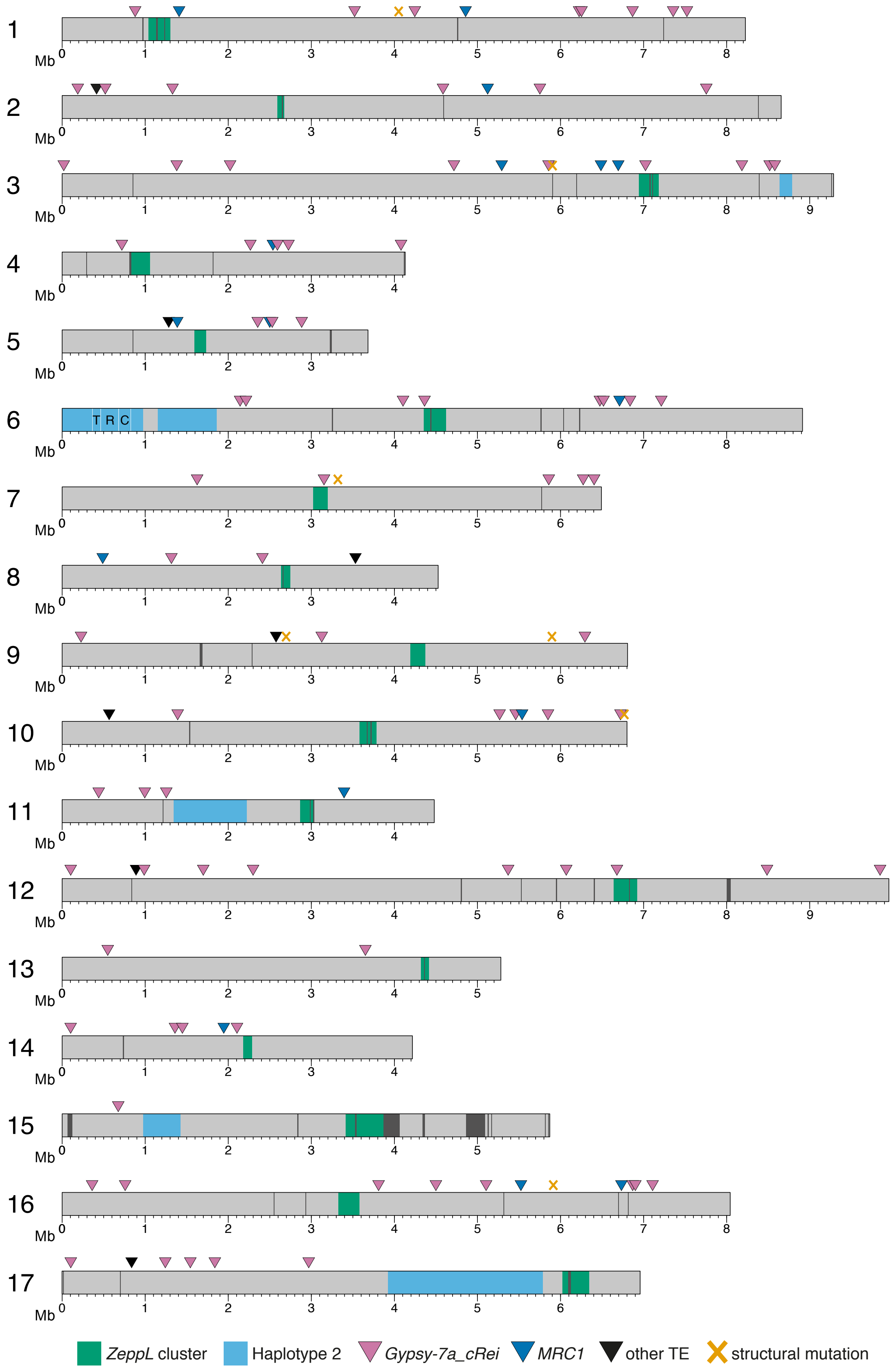
It therefore appears that the genome of CC-503 has experienced a substantial number of major structural mutations, potentially as a result of its mutagenesis (Hyams and Davies 1972). However, there is no evidence supporting the presence of the dupINVdup/deletion double mutation (Figure 4D) in v5, which appears to be entirely consistent with CC-1690/CC-4532 v6 across this region. Indeed, of the 74 structural mutations detected in CC-503 v6, 49 were unique to CC-503 v6, 19 were consistent between v5 and CC-503 v6, and six were ambiguous (Dataset S2). As described in 1.3.1, v5 is primarily based on the Sanger sequencing that was performed for the initial genome project in the early 2000s, with contributions from later Sanger and 454 sequencing. Conversely, the PacBio sequencing used in this study was performed on DNA extracted from a CC-503 culture maintained from stock obtained from the Chlamydomonas Resource Centre by Gallaher et al. (2015). Given the number of shared structural mutations, there appears little doubt that this more recently acquired culture shares a clonal common ancestor with the CC-503 used in the original genome project. Many of these shared mutations are large and distinctive, including a second 48.2 kb deletion on chromosome 16 and presumably also the reciprocal translocation (which was misassembled in v5 but still inconsistent with CC-1690/CC-4532 v6, Figure 4A). The most plausible explanation is therefore that the CC-503 genome is actively (and possibly increasingly) undergoing mutational degradation, with the majority of structural mutations having occurred in the laboratory over the preceding decade or so. The indels and duplications observed at DSBs and repair sites in Figure 4 (which were also observed flanking many of the other structural mutations) may be indicative of repair via an alternative non-homologous end joining (A-NHEJ) pathway, as opposed to the classical NHEJ (C-NHEJ) pathway that would not be expected to produce such large indels (Sfeir and Symington 2015; Chang et al. 2017). Although we did not observe any mutations in the genes encoding the core C-NHEJ machinery (*Ku70*, *Ku80*, *Artemis*, etc.), it is possible that CC-503 is in some way deficient in DSB repair. Finally, we found no clear gene candidates for the cell wall-less phenotype amongst the mutations consistent with v5, some of which presumably occurred under the original mutagenesis. It is possible that one of the affected genes is responsible via an unknown pathway, or that the phenotype is caused by a single nucleotide mutation or smaller indel.

4.4.5 A reference quality assembly of CC-4532 and a complete mating type *minus* locus

In the above sections we documented assembly improvements using the CC-503 v6 assembly, since it is based on the same strain as previous versions and simplifies a direct comparison. Given what is now known about structural mutations in the CC-503 genome, users may consider transitioning from v5 to our assembly of the *MT*$^{-}$ strain CC-4532 (i.e. CC-4532 v6), the relative merits of which are discussed in 4.4.9. Although the exact parent strains of CC-4532 are unknown (Gallaher et al. 2015), the strain is frequently used experimentally (e.g. Malasarn et al. (2013); Kropat et al. (2015)). We produced the CC-4532 v6 assembly using the same approach as for CC-503 v6, from a comparable PacBio dataset (4.5.1). The ordering of CC-4532 contigs was entirely consistent with CC-1690, and CC-4532 v6 was generally equivalent or better than CC-503 v6 based on comparison of assembly metrics (Table 1). CC-4532 v6 consists of 80 chromosomal and 40 unplaced contigs, with a contig N50 of 2.65 Mb and a total length of 114.0 Mb (Table S4). Slightly higher proportions of TEs (12.3%), microsatellites (1.8%) and satellite DNA (4.1%) were assembled relative to CC-503 v6, and as expected the 63 chromosomal gaps were generally associated with highly repetitive parts of the genome (Dataset S4). Unsurprisingly, chromosome 15 was the most fragmented (11 contigs, 9.2% gaps). The increased assembly size and repeat content could partially be attributed to the far more complete assembly of putative centromeres, with CC-4532 v6 containing 98.1% of *ZeppL* cluster sequence relative to CC-1690, and the putative centromeres of chromosomes 5, 7, 9, 14 and 16 appearing to be entirely assembled without gaps (Figure 6, Table S5). Subtelomeric and telomeric sequences were also better assembled relative to CC-503 v6, with 26 of the 34 CC-4532 v6 chromosome termini including telomeric repeats.

All laboratory strains of *C. reinhardtii* are thought to be the derived from the haploid progeny of a single diploid zygospore isolated by G. M. Smith in 1945 (1.2.3). Gallaher et al. (2015) showed that the genomes of laboratory strains are comprised of two haplotypes, with the haplotype not found in CC-503 (referred to as haplotype 2) present at up to ~25% of the genome in other strains. The two haplotypes differ at ~2% of sites, which is approximately equivalent to the genetic diversity observed between field isolates from the same location (Craig et al. 2019), consistent with the expectation that the haplotypes were inherited from a single zygospore. Relative to CC-503 (which is entirely haplotype 1 by definition, 1.2.3), CC-4532 contains six haplotype 2 regions spanning 5.0 Mb or 4.4% of the genome (Figure 5, Table S6). While the most conspicuous of these contains the *MT*$^{-}$ locus on the left arm of chromosome 6, the longest is a 1.9 Mb region on chromosome 17. Many variants found in haplotype 2 are predicted to alter protein sequences and are expected to underlie phenotypic differences between strains (Gallaher et al. 2015).

The ~461 kb *MT*$^{-}$ locus was entirely assembled on a single contig (Figure 5). Genes within the rearranged domain (R domain) are generally shared, but not syntenic, between *MT*$^{+}$ and *MT*$^{-}$ (and therefore CC-503 v6 and CC-4532 v6), although the two *MT*$^{-}$-specific genes (*MID* and *MTD1*) are obviously unique to CC-4532 v6. An *MT*$^{-}$ locus has previously been



1
Mb 0 1 2 3 4 5 6 7 8
2
Mb 0 1 2 3 4 5 6 7 8
3
Mb 0 1 2 3 4 5 6 7 8 9
4
Mb 0 1 2 3 4
5
Mb 0 1 2 3
6
T R C
Mb 0 1 2 3 4 5 6 7 8
7
Mb 0 1 2 3 4 5 6
8
Mb 0 1 2 3 4
9
Mb 0 1 2 3 4 5 6
10
Mb 0 1 2 3 4 5 6
11
Mb 0 1 2 3 4
12
Mb 0 1 2 3 4 5 6 7 8 9
13
Mb 0 1 2 3 4 5
14
Mb 0 1 2 3 4
15
Mb 0 1 2 3 4 5
16
Mb 0 1 2 3 4 5 6 7 8
17
Mb 0 1 2 3 4 5 6
ZeppL cluster
Haplotype 2
Gypsy-7a_cRei
MRC1
other TE
structural mutation

Figure 5. The CC-4532 v6 assembly.
Overview of the CC-4532 v6 chromosomes, showing putative centromeres (*ZeppL* clusters), haplotype 2 regions, assembly gaps (dark grey regions), TE insertions and structural mutations. All shown TEs are verified *de novo* insertions unique to CC-4532 that have occurred in laboratory culture. The three domains (T, R and C) of the *MT*$^-$ locus are marked by white lines within the haplotype 2 region of chromosome 6.

sequenced from the field isolate CC-2290 (=S1 D2) (Ferris et al. 2010; De Hoff et al. 2013), which is ~3% divergent from laboratory strains (Craig et al. 2019). The CC-4532 v6 R domain (~211 kb) was entirely syntenic with that of CC-2290 (~218 kb), although with the expected level of genetic divergence present throughout and substantial variation in the repetitive sequences present in noncoding regions. In contrast to *MT*$^-$, the *MT*$^+$ R domain contains several *MT*$^+$-specific genes, which except for *FUS1* and the "16 kb repeat" genes appear to have originated from autosomal insertions (De Hoff et al. 2013). All of these putative insertions (*MTP0428*, the MTA region and the SRL region) are shared between CC-503 v6 and CC-1690, indicating that they are common to all laboratory strains and are unrelated to the exceptional mutational landscape of CC-503. Indeed, the only major difference between the R domains of CC-503 v6 and CC-1690 was in the assembly of the "16 kb repeats", which resulted in two assembly gaps in CC-503 v6 (Dataset S1). The "16 kb repeats" refers to a complex ~160 kb region of 17.2 kb tandem duplications, which collectively contain multiple *MT*$^+$-specific copies of *EZY2*, *OTU2* and *INT1* (De Hoff et al. 2013). Overall, the availability of well-annotated (4.4.7), complete (*MT*$^-$, CC-4532 v6) or near complete (*MT*$^+$, CC-503 v6) mating type loci representing the two haplotypes present amongst laboratory strains is expected to be a major resource for the community.

4.4.6 The CC-4532 assembly reveals transposable element proliferation in the laboratory

The most remarkable feature of CC-4532 v6 is its length, which at 114.0 Mb is ~2.5 Mb longer than CC-503 v6. While this can partly be explained by the better assembly of putative centromeres and the deletion bias of mutational events in CC-503, most of the increase can be directly attributed to TE expansion within the laboratory. In our analysis of structural mutations (4.4.4), considering only events identified in regions where both CC-503 and CC-4532 were haplotype 1 (therefore limiting calls to mutations that have occurred in the laboratory), we identified 27 derived TE variants unique to CC-503 v6 (Dataset S5). As with other structural mutations detected in CC-503 v6, nine of the TE variants were absent from v5, indicating very recent TE activity. For CC-4532 v6, we identified 109 derived TE variants, all of which were insertions (Dataset S6). Remarkably, 86 of the insertions were of the same 15.4 kb *Gypsy* LTR element (*Gypsy-7a_cRei*, Figure S6), contributing ~1.3 Mb of novel sequence to CC-4532 v6 (Figure 5). This element has not been reported to be active amongst laboratory strains, and indeed no insertions of *Gypsy-7a_cRei* were detected in CC-503 v6, in which the element is present in just two full-length ancestral copies. Fortunately, only 11 of the 86 insertions were predicted to disrupt coding sequence (Dataset S6), with intergenic insertions observed 2.6x more frequently than expected by chance. This either implies that the element has some mechanism of targeted insertion, or that individuals experiencing genic insertions have been selected against in the laboratory. The *Gypsy-7a_cRei* Gag-Pol polyprotein contains a plant homeodomain (PHD) finger (Figure S6), which may be involved in targeted insertions (5.4.2). While a small number of the insertions were unusual (either containing less or more than one full copy of the element), we did not observe any novel solo LTRs, which are formed by ectopic recombination.

Given the proliferation of *Gypsy-7a_cRei* in CC-4532, we used whole-genome re-sequencing data to test whether the element is active in any other laboratory strains. Based on the ratio of mean read coverage across the internal region of *Gypsy-7a_cRei* and mean read coverage at non-repetitive sites genome-wide, it should be possible to estimate copy number across strains. Performing this analysis on CC-503 re-sequencing data resulted in a coverage ratio of 1.95 (in agreement with the expected ratio of 2), while the coverage ratio for CC-4532 was 50.8 (Figure 6). This is considerably lower than the number of copies in CC-4532 v6, potentially indicating that the proliferation of *Gypsy-7a_cRei* is ongoing, with tens of new insertions occurring between the re-sequencing performed by Gallaher et al. (2015) and the PacBio sequencing performed in this study. Several strains appeared to contain only the two ancestral copies (CC-1009, CC-1010 & CC-1690) or to show only modest increases in copy number (e.g. CC-124, ratio 3.9). In addition to CC-4532, two strains showed evidence for extreme proliferation of *Gypsy-7a_cRei*, CC-4533 (ratio 45.4) and CC-125 (ratio 68.8). CC-4533 has recently been used in the Chlamydomonas Library Project (CLiP), a large-scale mutant library used to uncover gene function (Li et al. 2019). The high ratio for CC-125 is surprising, given that CC-503 was derived from this strain. The proliferation in CC-125 has therefore presumably occurred since the isolation of CC-503 in the early 1970s, and it is perhaps even possible that there are maintained cultures of CC-125 in which *Gypsy-7a_cRei* has not been active.

The laboratory activity of *Gypsy-7a_cRei* provides an opportunity to estimate transposition rate. Unfortunately, since the origins of CC-4532 are uncertain it is not possible to estimate the number of generations over which *Gypsy-7a_cRei* may have been active in this strain. It is, however, possible to derive an estimate based on the well documented separation of CC-503 and CC-125 (see above). The CC-125 genome harbours 138 unique single nucleotide variants relative to CC-503 and all other laboratory strains (Gallaher et al. 2015), which presumably accumulated in the ~45 years between the strains being split and re-sequencing being performed. Based on a single nucleotide mutation (SNM) rate of 9.63×10^{-10} (Ness et al. 2015) and ~75% of the genome being callable using NGS data, CC-125 has undergone ~1,740 generations in this period. This equates to ~39 generations per year, which is a plausible figure given that stocks are generally maintained for months without cell division. Removing the two ancestral copies, the ~67 inferred insertions in CC-125 (Figure 6) result in an estimated transposition rate of 3.47×10^{-10} (assuming a genome size of 111 Mb, given that the copy number was coverage based). There are several caveats with this estimate. First, it assumes that *Gypsy-7a_cRei* has been active over a 45 year period, when its proliferation may have started many years after CC-125 and CC-503 were separated. Second, deleterious insertions may have been removed by selection during laboratory culture. Third, transposition rate would perhaps more meaningfully be calculated per active copy. It is unclear if any of the new insertions have themselves contributed to further proliferation, which would potentially result in the overall *Gypsy-7a_cRei* insertion rate varying considerably through time. Nonetheless, the estimate derived is comparable to a result obtained from mutation accumulation (MA) lines of *D. melanogaster*, in which the average insertion rate of active TE families was of the same order as the SNM rate (Adrion et al. 2017). As suggested in 1.5, performing similar analyses on *C. reinhardtii* MA lines would be the optimal approach to estimate transposition rates for the species.

Considering other TEs, the most active element was *MRC1*, which was responsible for 17 insertions in CC-503 v6 (Dataset S5) and 16 insertions in CC-4532 v6 (Figure 5, Dataset S6). *MRC1* was originally described as a non-autonomous LTR element (Kim et al. 2006), however we observed patterns of 5' truncation and tandem insertion (with the longest insertion containing five complete copies of the ~1 kb element) and we therefore re-classify *MRC1* as a non-autonomous *Penelope*-like element (5.4.3). *MRC1* insertions were also enriched in intergenic regions, occurring 2.0x more frequently than expected by chance. Neupert et al. (2020) also reported high activity of *MRC1* amongst laboratory strains, and it is likely that this element is generally one of the most active TEs in the laboratory. Four of the other TEs causing insertions have been described as active in the laboratory previously, namely one insertion each of the *hAT* DNA transposon *Gulliver* (Ferris 1989), the *Kyakuja* DNA transposon *Tcr1* (Schnell and Lefebvre 1993) and the *EnSpm* DNA transposon *Tcr3* (Wang et al. 1998), and three insertions of the non-autonomous DNA transposon *Bill* (Kim et al. 2006). A further eight active TEs have only previously been described in Repbase or in the updated TE library (5.4.1). Taken collectively, these results suggest that TE activity between laboratory strains can be highly heterogenous, with the potential for rapid TE proliferation to cause non-negligible increases in genome size and to disrupt genic sequence.

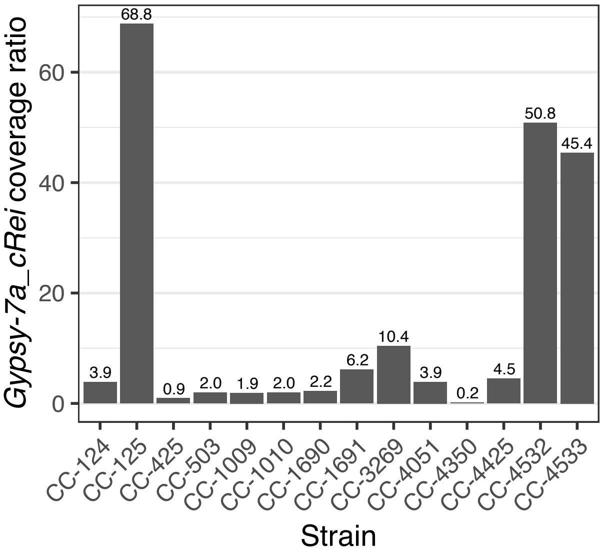


Figure 6. *Gypsy-7a_cRei* copy number estimates across laboratory strains.
Bars represent the ratio of whole-genome re-sequencing read coverage for the internal region of *Gypsy-7a_cRei* and non-repetitive genome-wide coverage.

4.4.7 Structural annotation of the version 6 assemblies

As with the assemblies, the initial structural annotations for CC-503 v6 and CC-4532 v6 were produced *de novo*. Data incorporated into the annotations included Iso-Seq, RNA-seq, and protein homology derived from the growing number of green algal gene annotations (4.5.5). To address the documented shortfalls in previous annotations (4.3), several additional steps were taken to augment the *de novo* annotations. As detailed in 4.5.6, efforts were made to ensure that no well-supported genes from previous versions were absent from the updated annotations. We also attempted to reduce the number of spurious gene models that showed little evidence of protein-coding capacity, and we accounted for a substantial number of genes that are likely part of TEs (4.4.8). We fixed the documented issue of the first in-frame start codon in a transcript regularly not being annotated as the start codon (Cross 2015). We manually curated twelve genes encoding selenoproteins (Novoselov et al. 2002), all of which were previously misannotated due to their use of the canonical stop codon "TGA" to encode selenocysteine. Although the gene models are finalised, work continues to update the gene IDs and to lift over and collate functional annotation data from previous versions and recent literature. The following section will briefly summarise the annotations and highlight improvements that have been made.

Table 2. Comparison of structural annotations between reference genome versions.

Annotation	Number of genes	Number of alternative transcripts	Transposable element genes	Low coding potential genes	BUSCO score (chlorophyta_odb10)
CC-503 v4.3	17,114	/	/	/	C:96.7%[S:96.0%,D:0.7%],F:1.3%,M:2.0%
CC-503 v5.6	17,741	1,789	/	/	C:98.9%[S:98.2%,D:0.7%],F:0.3%,M:0.8%
CC-503 v6.1	16,795	14,874	647	1,435	C:100.0%[S:99.3%,D:0.7%],F:0.1%,M:0.0%
CC-4532 v6.1	16,775	15,015	810	1,417	C:99.7%[S:98.7%,D:1.0%],F:0.1%,M:0.2%

The CC-503 v6.1 and CC-4532 v6.1 annotations contained 16,795 and 16,775 protein-coding genes, respectively (Table 2). Both of these figures were consistently lower than past versions (17,114 v4.3, 17,741 v5.6) as a result of the removal of low coding potential genes (see below) and TE genes (4.4.8). In contrast, the number of alternative transcripts increased considerably in both annotations. Supporting an overall improvement despite a lower gene number, the BUSCO (Benchmarking Universal Single-Copy Ortholog) scores were marginally higher for the updated annotations, with the number of fragmented and missing chlorophyte BUSCOs (n=1,519) dropping from five and eleven in v5.6, to one and zero (CC-503 v6.1) and one and three (CC-4532 v6.1) (Table 2). Of the three BUSCO genes that were present in CC-503 v6.1 but absent in CC-4532 v6.1, two could not be annotated due to assembly gaps in CC-4532, and the other was located in a haplotype 2 region. It was not clear why these regions failed to be assembled in CC-4532 v6, and although we expect almost all genic sequence to be assembled (4.4.3, 4.4.5), a very small number of genes may be affected by remaining gaps in CC-4532 v6. As RNA-seq data were not prepared specifically from each strain, haplotype differences between the underlying assemblies may also cause a small number of discrepancies between the two annotations. Although we attempted to lift over any genes present in one assembly but absent in the other, the high divergence (~2%) between the haplotypes meant that this was not always possible in regions where the strains differed in this respect.

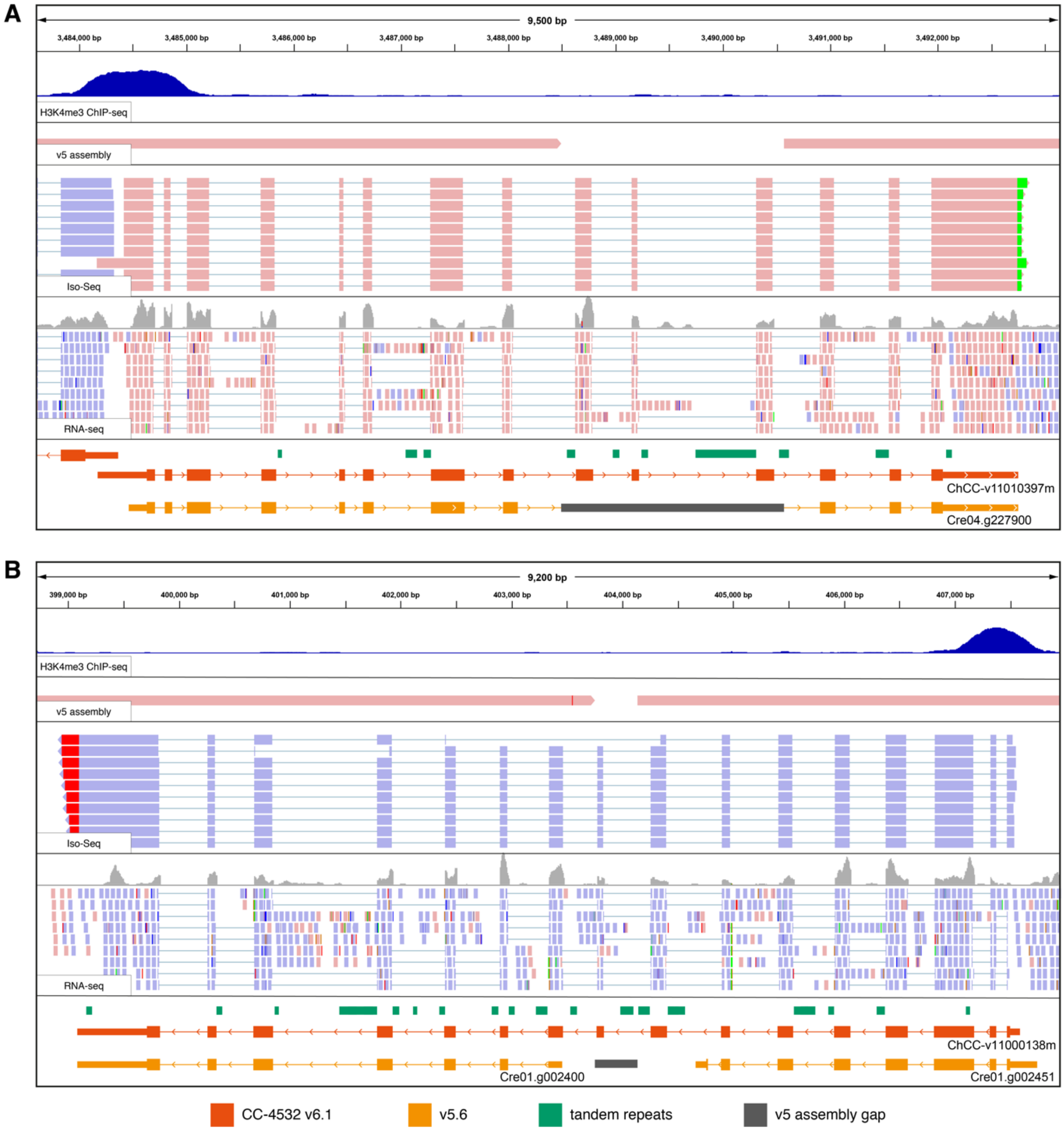


Figure 7. Browser view examples of gene models improved between v5.6 and CC-4532 v6.1. H3K4me3 ChIP-seq marks active promoters (Ngan et al. 2015). The v5 assembly track shows an alignment of v5 relative to CC-4532 v6, with assembly gaps appearing as unmapped regions (also see dark grey blocks). Exon coordinates for v5.6 gene models (orange) were lifted over to CC-4532 v6. Tandem repeats (green) includes both microsatellites and satellite DNA. Note that red and green mismatches at the end of Iso-Seq reads represent poly(A) tails.
(A) *PF20*, CC-4532 v6 coordinates: chromosome 4, 3,483,590 - 3,493,250.
(B) *TGL1*, CC-4532 v6 coordinates: chromosome 1, 398,620 - 407,978.

As introduced in 4.4.3, much of scope for improvement stems from the filling of assembly gaps in genic sequence. Two examples of improved genes are represented in Figure 7. Tulin and Cross (2016) highlighted *PF20*, which encodes a 606 amino acid protein involved in cilia function (Smith and Lefebvre 1997), as a gene with "hidden" exons in v5. The filling of the v5 assembly gap within *PF20* has resulted in the gene model being corrected, with three new exons (exons 9, 10 and 11) incorporated and the 3' splice site of exon 8 shifted (Figure 7A). While *PF20* was still annotated as a single gene in v5.6, other genes were fragmented by assembly gaps. *TGL1* is a putative triglyceride lipase orthologous to *HIL1* (*HEAT INDUCIBLE LIPASE1*) in *Arabidopsis thaliana*. In v5.6, the presence of an assembly gap resulted in *TGL1* being fragmented and annotated as two separate genes. The filling of this gap revealed a "hidden" exon, linking the two fragments as a single transcript (Figure 7B).

Finally, we introduced the concept of "low coding potential" genes in the v6 annotations. Craig et al. (2021a) used a combination of comparative genomics, population genetics and intrinsic features of protein-coding genes (codon usage bias and the strength of Kozak sequences) to quantify coding potential across the v5.6 gene models, reporting that several hundred genes were unlikely to be protein-coding. We repeated their analyses on the initial *de novo* annotations (4.5.5), from which we designated 1,435 (CC-503 v6.1) and 1,417 (CC-4532 v6.1) genes as low coding potential (Figures S7, S8). It is possible that many of these annotated transcripts are in fact long noncoding RNA genes that contain spurious ORFs long enough to be annotated by *ab initio* approaches. These genes will be made available as separate GFF3 (General Feature Format 3) files and we encourage genome users to explore these datasets where appropriate.

4.4.8 Transposable element genes and the interesting example of *CMD1*

Autonomous TEs contain genes necessary for their transposition, and certain TEs can also contain additional accessory genes. While these genes can have unusual features (e.g. ribosomal frameshifts) and may undergo frequent pseudogenisation, many are transcribed and they can be readily identified by gene prediction algorithms. The inclusion of TE genes within annotations can confound several analyses, including between species orthology and gene family analyses, and any study where substantial differences may be expected between standard genes and TEs (e.g. analyses of methylation or small RNA targeting). Most annotation projects therefore aim to filter as many TE genes as possible, while highly curated annotations for model organisms may include TE genes as defined entities. For the v6 structural annotations we applied a balanced approach, which aimed to incorporate a set of highly supported TE genes, while filtering lower confidence gene models that would require further curation (4.5.7).

Craig et al. (2021a) identified ~1,000 genes that overlapped the updated TE library (5.4.1) and are likely TE genes. The distribution of TE overlap for v5.6 genes was highly bi-modal, 1,023 genes had >30% CDS overlap with TEs, 908 of which had overlap >80% (Figure 8A). Similar distributions were also observed for genes in the initial v6 annotations, indicating that almost all predicted genes can be cleanly divided into TE and non-TE sets. To designate sets

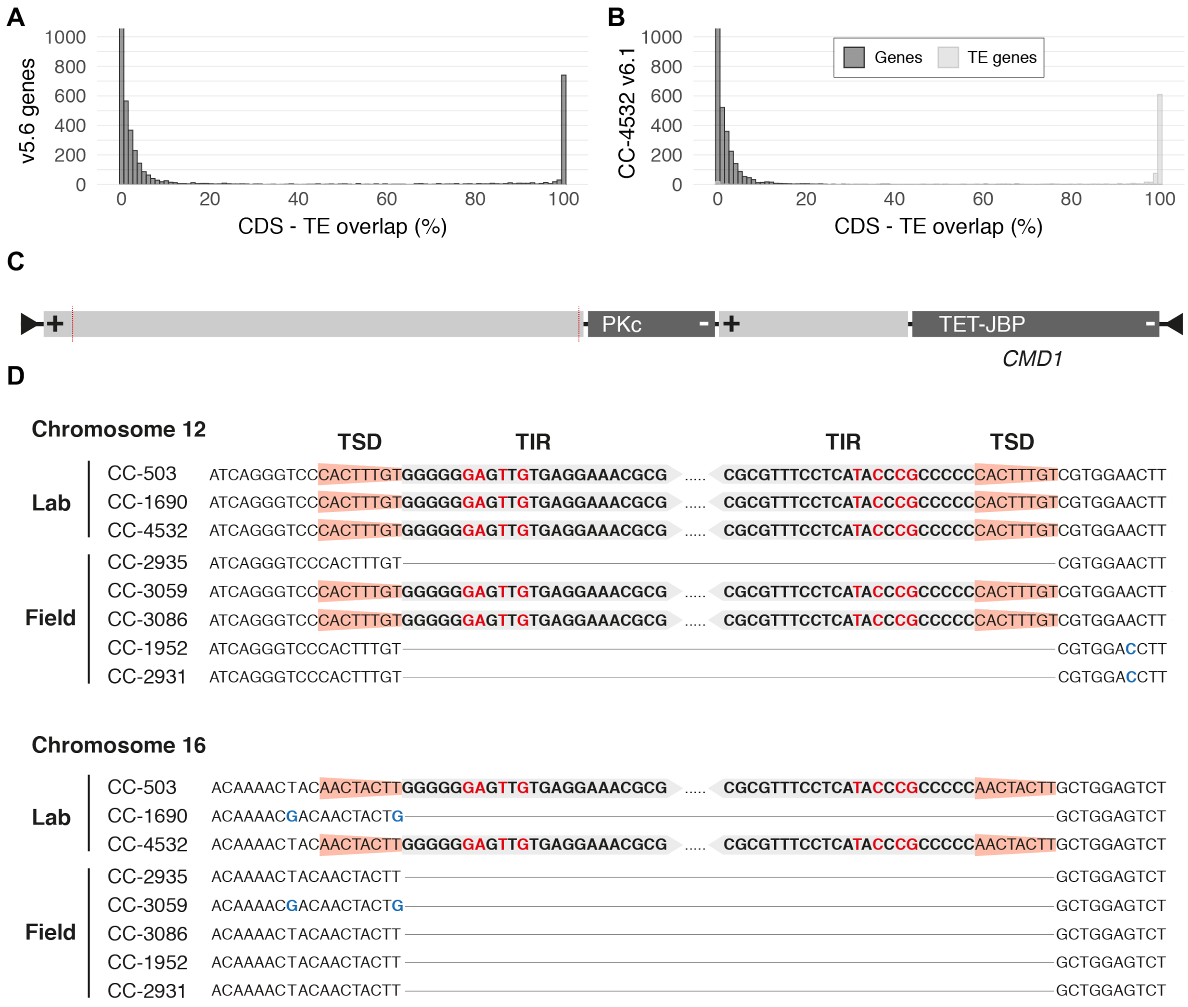


Figure 8. Transposable element genes and *CMD1*.
(A) Overlap between v5.6 per gene CDS and TEs. Note that not all genes are shown as >>1000 genes have 0% overlap.
(B) Overlap between CC-4532 v6.1 per gene CDS and the TE library. Genes are split into standard and TE genes.
(C) Schematic representation of the *DNA-1_cRei* transposon that contains *CMD1*. Gene strand is shown by +/- and domains are highlighted (PKc = protein kinase catalytic domain, TET-JBP = ten-eleven translocation/J-binding protein). The ~8 kb region deleted in the chromosome 12 copy is shown by the red dashed lines. Note that all genes have introns which are not shown.
(D) Sequence context of chromosome 12 and 16 insertions of *DNA-1_cRei*. The 8 bp target site duplications (TSDs) and the first 23 bp of the 67 bp terminal inverted repeats (TIRs) on each terminus are highlighted. Mismatches between the TIRs are shown in red text, and single nucleotide variants present in laboratory and field isolates are shown in blue text. Regions lacking the insertions are shown as solid black lines. Variants were manually curated from PacBio and Illumina data.

of high confidence TE genes, we required that genes with high repeat overlap had either a blastp hit to a known TE protein and/or a functional domain. This resulted in the inclusion of 647 TE genes in CC-503 v6.1 (Table 2, Figure S9) and 810 in CC-4532 v6.1 (Figure 8B), which are included in the GFF3 files under the type field “transposable_element_gene”. Users should be aware that these are not exhaustive TE gene sets, and that >4,000 preliminary gene models were entirely filtered from each annotation based on repeat overlap.

Projects requiring coordinates of TEs in general should use dedicated repeat annotation tracks.

Most of the TE genes encode proteins with typical TE domains, such as reverse transcriptases, endonucleases and transposases. These genes can be included or excluded in analyses as desired (for example including the gene models in annotation-guided RNA-seq mapping may improve results). However, it is important to note that TE proteins are capable of important and wide ranging interactions with the host genomic and cellular environments (Cosby et al. 2019) and that the TE gene sets are also of functional relevance. A specific example of this in *C. reinhardtii* is the recently described TET/JBP (ten-eleven translocation/J-binding protein) gene *CMD1* (5mC-modifying enzyme). Xue et al. (2019) demonstrated that *CMD1* is responsible for DNA demethylation via a novel biochemical pathway, converting C^5-methylcytosine (5mC) to C^5-glyceryl-methylcytosine (5gmC) using a vitamin C co-substrate. Genome-wide 5mC levels were doubled in a *cmd1* mutant, which resulted in downregulated expression of certain genes and a susceptibility to photodamage (Xue et al. 2019). However, *CMD1* is not a host gene, and appears to be part of a DNA transposon (annotated as *DNA-1_cRei* in the updated library, 5.4.1). The highly unusual 18.4 kb TE encodes four genes (including *CMD1*), although none are a recognised transposase (Figure 8C). *DNA-1_cRei* nonetheless possesses all other characteristics of a DNA transposon, namely 67 bp imperfect terminal inverted repeats, precise insertion polymorphisms between both laboratory strains and field isolates, and 8 bp target site duplications (Figure 8D). In CC-503 v6 and CC-4532 v6, *DNA-1_cRei* is present in two copies, one on chromosome 12 that contains an ~8 kb internal deletion, and a second presumably full-length copy on chromosome 16. CC-1690 lacks the chromosome 16 copy (Figure 8D), although CC-1690 is haplotype 2 in this region and the variant is presumably ancestral and not related to activity within the laboratory. TET-JBP genes have been described in *KDZ* DNA transposons in fungi, where their DNA demethylation activity is thought to play a role in self-regulation and the regulation of other TEs (Iyer et al. 2014). Furthermore, *C. reinhardtii* contains at least ten TET/JBP genes in addition to the one or two copies of *CMD1* (Aravind et al. 2019), most of which appear to be carried by copies of two Helitron families (5.4.1). With the exception of hypermethylated regions (4.4.3), *C. reinhardtii* has very low levels of 5mC methylation (<1% genome-wide) (Lopez et al. 2015), and it is possible that the presence of multiple TEs containing TET/JBP genes is partly responsible for this. This interesting case demonstrates that TE genes can underlie important biological processes and are capable of producing notable phenotypes. Such genes can vary in copy number between laboratory strains, and their inclusion within the wider annotation should facilitate the study of other novel aspects of *C. reinhardtii* biology.

4.4.9 The present and future of the Chlamydomonas Genome Project

Over nearly the last two decades the *C. reinhardtii* reference genome has been based on a single strain, CC-503. It has long been known that *C. reinhardtii* laboratory strains differ both genetically and phenotypically, and it has been demonstrated that most of this variation stems from the two haplotypes present among strains (Gallaher et al. 2015). These developments have led to the "know thy strain" maxim (Salomé and Merchant 2019), with researches encouraged to consider the genetic differences that may exist between the reference genome and other strains used in experimental work. Perhaps the most revealing finding of this study is that these differences should not only be considered with respect to ancestral variation between the two haplotypes, but also to derived variation arising by mutation in the laboratory. Although the large number of structural mutations in the mutagenised CC-503 genome represents an extreme case, the CC-4532 genome harbours seven structural mutations and more than 100 TE insertions. Indeed, both Flowers et al. (2015) and Gallaher et al. (2015) used short-read data to identify several structural variants segregating in haplotype 1 regions among laboratory strains, including major putative duplications spanning hundreds of kb. While it was already known that certain phenotypes were caused by laboratory mutations (e.g. *nit1* and *nit2*), it is likely that all strains have experienced unique structural mutations, and that certain strains are experiencing TE proliferation. It is also possible that maintained cultures of the same strain may differ due to mutation. As demonstrated with the CC-503 v6 and CC-4532 v6 assemblies, a proportion of these mutations and TE variants are expected to disrupt genes. Many laboratory strains have been maintained clonally for approaching 75 years and mutations are of course an unavoidable consequence of this, especially given that strains are likely evolving under relaxed selection in laboratory conditions. Some mutant phenotypes are even desirable in certain circumstances, for example cell wall-less mutants have been frequently used for gene transformation. The implications of "laboratory domestication" have been considered in other laboratory systems such as *Caenorhabditis elegans* (Sterken et al. 2015), and laboratory mutations should be carefully considered when interpreting experimental results in future *C. reinhardtii* research.

Based on the results discussed above, many genome users are likely considering which assembly to use for their future research. Both the CC-503 v6 and CC-4532 v6 assemblies are highly contiguous, biologically accurate and well-annotated. Although the CC-503 genome has been shown to contain large structural mutations, CC-503 v6 will nonetheless be an excellent resource for the community. The mutations affect <1% of genes, and the assembly and annotation are a substantial improvement on v5. The availability of an annotated MT^{+} locus in the CC-503 v6, which does not appear to have been affected by mutations, is expected to be particularly useful. Nonetheless, since CC-4532 v6 is comparable on all metrics to CC-503 v6, on balance it will be the best assembly for most projects. While the CC-4532 genome has experienced many TE insertions, the majority of these are intergenic and considerably fewer genes have been disrupted relative to CC-503 v6. CC-4532 v6 is also the obvious choice for any analyses that require a "wild type" karyotype (e.g. recombination studies). Whichever assembly researchers decide to use as a primary reference, the availability of two assemblies and annotations of such high quality provides the major benefit that users can assess loci of interest in CC-503 v6 and CC-4532 v6, ensuring that their results are supported by both.

Considering the future of the genome project, it will likely soon be possible to produce complete *C. reinhardtii* reference assemblies using existing technologies. However, given that the assembly gaps in CC-503 v6 and CC-4532 v6 are generally limited to highly repetitive regions, we expect that almost all genic sequence is now assembled, and the necessity for any further improvements is largely reduced to very specific analyses involving regions such as centromeres or subtelomeres (which could already be performed on the CC-1690 assembly). We therefore expect that most improvements will be concentrated on structural annotation. Although v6 annotations are undoubtedly the most high-quality to date, their remains scope for improvement. The availability of various "omics" data for *C. reinhardtii* and related species continues to grow, presenting opportunities to integrate new annotation evidence. This may be especially important for the annotation and analyses of alternative isoforms and polycistronic loci, which have not been considered in this study. The main annotations also currently lack small and long noncoding RNA gene models, and there are substantial opportunities to enhance annotations beyond protein-coding genes.

With the availability of reference-quality assemblies and annotations for three and two strains, respectively, we are now entering an exiting new era of *Chlamydomonas* genomics. As outlined above, when considering alternative haplotypes and derived mutations, no single reference assembly is expected to be optimal for all laboratory strains. One can now imagine the future development of a *C. reinhardtii* pan-genome, where assemblies from several strains would enable all haplotype 1 and 2 regions to be represented and used as appropriate for a strain of interest. Such approaches are already possible in a limited sense, for example a project using both MT^{+} and MT^{-} strains would be encouraged to use a custom reference including both loci (e.g. see Ness et al. (2015)). With data from several strains, it may even be desirable to produce consensus assemblies for each haplotype, inferring the ancestral state at the time of isolation and accounting for derived mutations. Furthermore, there is far greater diversity present amongst *C. reinhardtii* field isolates (Flowers et al. 2015; Craig et al. 2019), which could be incorporated into a true species-level pan-genome. Such developments are expected to reveal novel aspects of *C. reinhardtii* biology, continuing the development of the species as an integral model in plant and algal biology.

4.5 Methods

4.5.1 Chromosomal assemblies of CC-503 and CC-4532

Both CC-503 v6 and CC-4532 v6 were assembled using similar approaches. DNA extraction, sequencing and the assembly of preliminary contig-level assemblies were performed by others and will be described in detail elsewhere. Briefly, the initial CC-503 assembly was produced with MECAT (Xiao et al. 2017) from 127x PacBio Sequel reads with a mean read length of 3,577 bp. The initial CC-4532 assembly was produced with Canu (Koren et al. 2017) from 176x coverage of PacBio Sequel reads with a mean read length of 9,883 bp. Both assemblies were subsequently polished using the PacBio reads and Arrow (https://github.com/PacificBiosciences/GenomicConsensus), followed by additional error correction using >50x of Illumina data.

To scaffold the resultant contigs to chromosomes, we primarily relied on ordering and orientating the contigs relative to the CC-1690 assembly. Contigs were mapped to CC-1690 using two tools, MashMap v2.0 (Jain et al. 2018a) with the parameters "--perc_identity 95" and "-f one-to-one", and minimap2 v2.17 (Li 2018) with the parameter "-ax asm5". MashMap is useful for interpreting broad-scale mapping since it does not explicitly perform alignment between input sequences, instead identifying approximate alignment boundaries using a kmer-based similarity between sequences (with the minimum reported segment set to 5 kb by default). Conversely, minimap2 performs explicit alignment, providing fine-scale mapping and precise genomic coordinates for mapped sequences. The resultant contig vs chromosome maps were inspected manually, and in almost all cases both tools produced consistent results at the scale of entire contigs. Any inconsistencies between the contigs and chromosomes (i.e. a contig not mapping consistently to a single chromosomal region or not mapping with a consistent orientation) were validated by manually inspecting the relevant PacBio reads mapped to that region with IGV v2.7.2 (Robinson et al. 2011). In a small number of cases this resulted in a misassembled contig being split, while for the CC-503 contigs several inconsistencies were clearly supported by the reads and were assembled as structural mutations (4.4.4). Contigs that did not contain any uniquely mapping regions were retained as unassembled contigs. Contigs consisting of entirely subtelomeric repeats, which generally did not map uniquely, were assigned to chromosome termini by manual comparison (performed by Olivier Vallon) to the satellite repeats of the subtelomeres of CC-1690 (Chaux-Jukic et al. 2021). The curated contig to chromosome maps were then validated against the molecular markers of Kathir et al. (2003), the coordinates of which were located in the assemblies by megablast.

Gap lengths between contigs were estimated relative to the assembled sequence in CC-1690 (using the mapping coordinates of minimap2) and the appropriate number of Ns were inserted between contigs. In a number of cases the calculated gap was negative, suggesting redundant sequence either side of the gap. These contig ends were compared, redundant sequence was trimmed, and the two contigs were merged if possible (performed by Jerry Jenkins). Arbitrary gaps of 100 Ns were inserted between contigs in cases which could not be successfully merged.

4.5.2 Identification of repetitive sequences

TE sequences were identified in each genome by providing the updated *C. reinhardtii* TE library (5.4.1) to RepeatMasker v4.0.9 (Smit et al. 2013-2015). Any TE sequences >20% divergent from their respective consensus sequence were removed. *ZeppL* clusters (i.e. putative centromeres) were identified as the span from the first two consecutive *ZeppL-1_cRei* copies to the final two consecutive *ZeppL-1_cRei* copies on each chromosome.

Microsatellites and satellite DNA were identified using Tandem Repeats Finder (Benson 1999) with the recommended parameters "2 7 7 80 10 50 500" (i.e. a minimum alignment score of 50 and a maximum period size of 500 bp). Tandem repeats of >2 copies were split to microsatellites (if the monomer was <10 bp) and satellite DNA (monomers >10 bp). If a given region was called as both a microsatellite and satellite DNA, satellite DNA was given preference since shorter monomers are frequently identified within larger ones. Subtelomeric satellites, which have monomers longer than the 500 bp identified by Tandem Repeats Finder, were identified by megablast using the sequences from Chaux-Jukic et al. (2021).

4.5.3 Comparison and lift over between assemblies

To determine genomic regions affected by misassembly corrections, the v5 assembly was mapped against CC-503 v6 using MashMap (4.5.1). The genomic coordinates of intra- and inter-chromosomal inconsistencies were then assessed manually and converted to input files for Circos (Krzywinski et al. 2009) and karyoploteR (Gel and Serra 2017) to produce Figures 1, 2 and S1. The hypermethylated regions identified in the v5 assembly by Lopez et al. (2015) were also mapped forward using MashMap.

Precise lift over of genomic coordinates between assemblies was performed via the production of a 5-way Cactus whole-genome alignment (WGA) (Armstrong et al. 2019) of the v4, v5, CC-503 v6, CC-4532 v6 and CC-1690 assemblies. Genomes were first soft masked for repeats by providing the genomic coordinates of TEs, microsatellites and satellite DNA (4.5.2) to the BEDtools v2.26.0 (Quinlan and Hall 2010) tool maskfasta, run with the option "-soft". An arbitrary guide tree for Cactus was provided as "(CC-4532_v6:0.001,(CC-1690:0.001,(CC-503_v4:0.001,(CC-503_v5:0.001,CC-503_v6:0.001):0.001):0.001):0.001)", and all genomes were selected as reference quality. Lift over of genomic coordinates between any pair from the five assemblies in the WGA was then performed using the HAL Tools command halLiftover (Hickey et al. 2013).

To identify filled gaps in the v5 assembly, the 250 bp sequences flanking either side of each gap were lifted over to the CC-503 v6 assembly. A gap was called as filled if the lift over identified at least 50 bp of sequence exhibiting one-to-one alignment on each flank, and the intervening sequence between the flanks in CC-503 v6 did not contain Ns. 1,012 gaps were unambiguously called as filled, 306 of which had negative lengths (suggesting redundant sequence either side of the gap in v5) and were not analysed further. The remaining 706 filled gaps were classified based on their repeat content. Filled gaps that contained >50% of either TEs, microsatellites or satellite DNA were classified to those respective categories, gaps that contained >25% repeats but no overall majority were classified as "repetitive", and all other gaps were classified as "other".

4.5.4 Curation of structural mutations and transposition events

Structural mutations were identified using the tool MUM&CO (O'Donnell and Fischer 2020), which calls variants from alignments produced by MUMmer (Kurtz et al. 2004). MUM&CO was run between CC-503 v6 and CC-4532 v6 on each of the 17 chromosomes individually. For chromosomes 2 and 9, the CC-503 v6 chromosomes were split at the translocation breakpoints and the relevant parts of each chromosome were included. All called variants were then manually curated by comparing the CC-503 v6, CC-4532 v6 and CC-1690 assemblies in IGV (using alignments produced by minimap2, 4.5.1). Variants called within tandem repeats and within regions where CC-4532 was haplotype 2 were not considered. Mutations were polarised by comparison of the three assemblies i.e. the allele present in two assemblies (one of the v6 assemblies and CC-1690) was assumed to be ancestral.

Transpositions were special cases of insertion and deletion events called by MUM&CO. Using IGV, if an insertion or deletion perfectly corresponded to the genomic coordinates of a TE (4.5.2) then it was called as a transposition. Most cases were insertions (i.e. one genome contained an insertion variant relative to the two other genomes), although a small number of DNA transposon excisions were also called (i.e. one genome contained a deletion).

To estimate the copy number of *Gypsy-7a_cRei*, the CC-503 v6 assembly was hardmasked for all TEs. The hardmasked assembly was then concatenated with the repeat library, in which the *Gypsy-7a_cRei* consensus is split into its component LTR and internal parts. Illumina reads from various laboratory strains sequenced by Gallaher et al. (2015) were then mapped to the assembly/repeat library, and the ratio between the mean coverage for the *Gypsy-7a_cRei* internal region and mean genome-wide coverage was calculated (performed by Sean Gallaher).

4.5.5 Structural annotation and designation of low coding potential genes

De novo annotations for CC-503 v6 and CC-4532 v6 were performed by Shengqiang Shu and will be described in detail elsewhere. Briefly, approximately 290,000 transcript models were build using PASA (Haas et al. 2003) from ~1.6 billion 150 bp paired-end RNA-seq reads, ~520 million 50 bp unpaired RNA-seq reads, ~6.4 million 454-sequenced ESTs and ~1.6 million Iso-Seq reads. Collectively, the transcriptomic sequencing was performed on several different laboratory strains. Gene models were predicted using four tools: either directly from the PASA transcripts by identifying ORFs, or using FGENESH+/FGENESH_EST (Salamov and Solovyev 2000), AUGUSTUS (Stanke et al. 2006; Stanke et al. 2008) or EXONERATE (Slater and Birney 2005). Where possible, protein homology was incorporated from 13 green algal annotations and a further nine plant and animal annotations. The best scoring gene at a given locus was retained, with the score based on the agreement with the transcriptomic data and protein homology, as well as overlap with repeats. PASA was used to improve the retained gene models by adding UTR sequence and alternative transcripts.

In order to minimise the number of false positive gene models included, the coding potential analyses of Craig et al. (2021a) were repeated. Briefly, for each annotation all genes were split into two classes, either control (genes with at least one algal homolog or a recognised protein domain) or test (genes with neither a homolog nor a protein domain). Algal homologs and protein domains were identified using OrthoFinder v2.2.7 (Emms and Kelly 2015) and InterProScan v5.39-77.0 (Jones et al. 2014), respectively, as performed by Craig et al. (2021a). Four metrics were then calculated for each gene, the PhyloCSF score (Lin et al. 2011), the ratio of genetic diversity at zero-fold and four-fold degenerate sites ($\pi_{0D/4D}$), codon usage as quantified by the index of translation elongation (I_{TE}) (Xia 2015), and a Kozak score that quantifies the strength of the Kozak sequence (Cross 2015). Two 8-way Cactus WGAs were produced, one with CC-503 v6 plus seven core-*Reinhardtinia* genomes (*Chlamydomonas incerta*, *Chlamydomonas schloesseri*, *Edaphochlamys debaryana*, *Gonium pectorale*, *Yamagishiella unicocca*, *Eudorina* sp. and *V. carteri*) and one with CC-4532 v6 plus the same seven genomes. The guide tree was identical to that used by Craig et al. (2021a), as was post-processing of the WGAs and the subsequent calculation of the PhyloCSF scores. Genetic diversity was estimated for both CC-503 v6.1 and CC-4542 v6.1 genes using whole-genome re-sequencing data from 17 Quebec field isolates, with single nucleotide polymorphisms called and filtered for both v6 assemblies following Craig et al. (2019). I_{TE} was calculated using the optimal codon usage table previously produced using v5.6 genes (Craig et al. 2021a). Kozak scores were calculated as described previously. Briefly, a Kozak sequence consensus logo, which summarises the frequency of nucleotides 5 bp up- and downstream of start codons, was produced using a random half of the control set genes. The start codons of each gene in the test set and the other half of the control set were then scored for their match to the consensus logo, with a higher score indicating a stronger Kozak sequence. A test set gene was determined to be low coding potential if it had a PhyloCSF score <1 and failed at least two of the following conditions: $\pi_{0D/4D}$ >95th percentile of $\pi_{0D/4D}$ for the control set genes, I_{TE} <5th percentile of I_{TE} for control set genes, or a Kozak score <0.25. Low coding potential genes with three or more exons, or ORFs ≥900 bp (after subtracting tandem repeats), were retained in the main gene sets.

4.5.6 Lift over of missing genes between annotation versions

To avoid any well-supported genes from previous annotations failing to be included in the v6 annotations, we attempted to lift over any missing genes. We started with a dataset of genes from the v4 assembly that were absent in the v5 annotations (Blaby and Blaby-Haas 2017), the entire v5.6 gene set, and the 142 high confidence genes identified by Craig et al. (2021a) that were missing from v5.6. Any genes with CDS having >30% overlap with TEs were removed and all genes were required to have strong evidence of coding potential (either an algal homolog, a recognised protein domain or a PhyloCSF score ≥100). The CDS of each gene was then lifted over from the relevant source genome to the relevant target genome using halLiftover (4.5.3). Genes were considered as potentially missing in the target genome if ≥90% of CDS sites could be lifted over, and if <10% of the lifted over sites overlapped existing CDS in the target genome. Additionally, both v6 annotations were checked against each other i.e. all CC-503 v6.1 genes with strong coding potential were lifted over to CC-4532 v6 and checked for overlap with the CC-4532 v6.1 genes, and vice versa. In cases of redundancy between v6.1 and v5.6 genes, v6.1 genes were given preference.

We then attempted to add the preliminary datasets of missing genes to the relevant v6 annotation using three approaches. If the CDS lift over was entirely one-to-one, then the gene was simply incorporated based on the lift over coordinates. If not (i.e. ≥90% of CDS was lifted over but <100%), we searched for the gene in question amongst the discarded low-scoring gene models from the *de novo* annotations (4.5.5). A match for a missing gene was considered to have been found if there was a significant blastp hit between the source and target predicted proteins (e-value ≤0.01 and percent identity ≥95%) and there was a ≥80% intersect between the CDS coordinates of the gene in the target genome and the lifted over CDS of the missing gene. If neither approach was successful, the transcripts of the missing genes were mapped as cDNAs to the relevant target genome with GMAP (Wu and Watanabe 2005) with the parameters "--cross-species --max-intronlength-ends 200000 -z sense_force". Gene models were included if the mapped cDNA had a valid ORF in the target genome.

4.5.7 Annotation of transposable element genes

In the *de novo* annotation (4.5.5), any genes with >20% CDS overlap by TEs or with >30% or their identified protein domains being TE-associated were filtered out, unless they had very strong homology support from proteins of other species, in which case the threshold for TE-CDS overlap was increased to 80% (note that many other species will have TE genes annotated in their main annotations, so strong homology is to be expected). This preliminary set of TE genes was then further analysed to produce high confidence sets of TE gene models. Each gene was queried against a database of TE proteins obtained from Repbase using blastp. A gene was considered to be high confidence if it either had any protein domain (i.e. not only TE-associated domains) or a significant blastp hit. A significant hit was defined as one with an e-value ≤0.001 and both query and hit spanning ≥50% of their respective lengths if the percent identity between the proteins was ≥60%, or spanning ≥20% of their respective lengths if the percent identity was <60%. All high confidence TE genes were manually reduced to a single isoform that had the highest support from the transcriptomic data. TE genes that did not pass the criteria were excluded.

Chapter 5

Transposable Element Annotation in *Chlamydomonas* Reveals a Major New Clade of Retrotransposons

5.1 Preface

The work in this chapter is almost entirely my own and the first-person singular is used throughout. After presenting this work at a conference I learnt that Irina Arkhipova had independently discovered a similar result, and some of the interpretations presented in this chapter were developed by useful discussions with the Arkhipova group. All data presented is my own, with the exception of the identification of hammerhead ribozymes which was performed by Fernando Rodriguez.

Excerpts of this chapter contributed to the following manuscript:

Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. 2021. An ancient clade of *Penelope*-like retroelements with permuted domains is present in the green lineage and protists, and dominates many invertebrate genome. *Mol Biol Evol* **in press**. doi:10.1093/molbev/msab225

5.2 Abstract

Transposable element (TE) annotations are a fundamental resource for many genomics analyses. The majority of TE annotation effort has been focussed on animal and plant genomes, and the curation of TEs in phylogenetically diverse species has often led to the discovery of entirely new types of TEs. Via exhaustive manual curation, I produce a *Chlamydomonas reinhardtii* TE library that greatly extends previous annotations and is expected to be near complete. Based on the newly described TEs in *C. reinhardtii* and its close relatives, I describe a novel clade of *Penelope*-like elements (PLEs). PLEs are an enigmatic clade of retrotransposons that share a common ancestor with telomerase reverse transcriptases (TERTs). Described PLEs form two major groups, telomere-restricted endonuclease deficient (EN–) elements that are present in several eukaryotic kingdoms, and elements containing a C-terminal GIY-YIG endonuclease (EN+) that are present in animals and transpose genome-wide. The PLEs discovered in *Chlamydomonas*, which are part of a major new group consisting of two superfamilies called *Chlamys* and *Naiad*, contain an N-terminal GIY-YIG endonuclease and exhibit patterns of genome-wide transposition. I functionally characterise an active *Chlamys* element in *C. reinhardtii*, demonstrating that these newly discovered elements likely share a mechanism with canonical EN+ PLEs. I search for other N-terminal EN+ PLEs across eukaryota, curating elements in more than 30 species of green algae, plants, animals and protists. I describe the first reported examples of TE-encoded selenoproteins in two animal *Naiad* elements. Phylogenetic analysis of the reverse transcriptase (RT) protein domain shows that the N-terminal EN+ elements form a clade basal to all other PLEs. Furthermore, the presence of an additional RT domain in *Chlamys* and *Naiad* that was previously only known from TERTs strengthens the evolutionary relationship between these major clades. These results imply more than one gain of genome-wide retrotransposition in PLEs and increase our understanding of the evolution of PLEs and TERTs in early eukaryotic genomes.

5.3 Introduction

As reviewed in 1.4.3, TEs are selfish mobile units of DNA that display an extraordinary diversity, are present in almost all eukaryotic genomes and have been implicated in a wide variety of biological phenomena. Thoroughly curated TE annotations are required to perform any detailed analyses of TEs in a species of interest. Furthermore, as demonstrated in Chapters 3 and 4, TE annotations are powerful resources for characterising general features of genome architecture and improving gene annotations.

C. reinhardtii has previously been the focus of significant TE annotation effort (1.3.7). Existing annotations include both experimentally characterised TEs (e.g. *TOC1* (Day et al. 1988) and *Gulliver* (Ferris 1989)) and TE families directly curated from early versions of the genome assembly (e.g. *Dualen* LINE elements (Kojima and Fujiwara 2005)) and *Novosib* DNA transposons (Kapitonov and Jurka 2008)). Most of these TEs are represented in a library of 119 consensus sequences available from Repbase (https://www.girinst.org/repbase/), which is generally used in *C. reinhardtii* analyses requiring TE annotations (e.g. Zhao et al. (2007); Philippsen et al. (2016)). However, there have been no attempts to update TE annotations in line with improvements to the reference genome and the completeness of the Repbase library is unclear. It is expected that the library will be biased towards TEs that were sufficiently assembled in the early assembly versions used for annotation (i.e. v3 and earlier, 1.3.1), which contained many more assembly gaps relative to v4 and onwards (Blaby et al. 2014). Furthermore, knowledge of TEs has progressed considerably in the past two decades and there are several examples of misclassified TEs in the Repbase library (5.4.1). Finally, annotations for other green algae are almost entirely restricted to an incomplete TE library for *Volvox carteri* (Lindauer et al. 1993; Miller et al. 1993; Duncan et al. 2002; Prochnik et al. 2010), and as a result, analyses in *C. reinhardtii* cannot rely on annotations from closely related species. Thus, a systematic review and update to the *C. reinhardtii* TE library is necessary.

Beyond the general uses for TE annotations in genomics analyses, the curation of TEs in taxa distantly related to typical model species has often led to the discovery of new types of TEs or increased our understanding of enigmatic TE groups (Wuitschick et al. 2002; Goodwin et al. 2003; Kojima and Fujiwara 2005; Böhne et al. 2012; Bao and Jurka 2013; Iyer et al. 2014; Ribeiro et al. 2019). One such group are PLE retrotransposons. The first described PLE, the active element *Penelope*, was discovered in *Drosophila virilis*, where its mutagenic activity is responsible for hybrid dysgenesis (Evgen'ev et al. 1997). As early genome and EST sequences were made available, it became apparent that *Penelope* represented a novel order of retrotransposons, with related elements discovered in several vertebrates (e.g. the fish *Fugu rubripes, Tetraodon nigroviridis* and *Oryzias latipes*, and the frog *Xenopus laevis*) and invertebrates (e.g. the flatworm *Schistosoma mansonii*, nematode *Anclyostoma caninum* and sea urchin *Strongylocentrotus purpuratus*) (Lyozin et al. 2001; Volff et al. 2001). Several structural and mechanistic features differentiated this new group from LINEs and LTRs, the two established retrotransposon orders. PLEs generally contain a single ORF encoding a unique configuration of RT and a C-terminal GIY-YIG EN, a type of nuclease originally discovered in homing endonucleases (Aravind et al. 1999; Kowalski et al. 1999). Phylogenetic analysis of the RT domain confirmed that PLEs are evolutionarily distinct from LINEs and LTRs, and instead form a sister clade to TERTs (Arkhipova et al. 2003), the non-mobile RT enzymes responsible for telomere maintenance in most eukaryotes. The GIY-YIG EN domain present in PLEs was also found to be highly divergent, containing an extremely conserved CCHH zinc finger motif within the core GIY-YIG domain that is found in no other GIY-YIG ENs (Volff et al. 2001; Arkhipova 2006). In a protein clustering analysis, the GIY-YIG ENs with highest similarity to those of PLEs were found in homing endonucleases of the Chlorovirus *Paramecium bursaria* chlorella virus 1 (PBCV-1) and the Iridovirus Chilo iridescent virus (CIV), and in the *Tlr8* Maverick/Polinton element of *Tetrahymena thermophila* (Dunin-Horkawicz et al. 2006).

PLEs are further characterised by their unusual ability to retain introns and their partial-tandem insertion, most frequently in a head-to-tail orientation (Evgen'ev and Arkhipova 2005). As PLEs undergo 5’ truncation, a complete insertion is represented by one full-length copy and a second partial upstream copy consisting of a variable length of the 3’ end, forming a pseudo LTR (pLTR). Functional characterisation of *Penelope* revealed that tandem insertion is a functional requirement, since the promoter and transcription start site (TSS) are located in the pLTR (Schostak et al. 2008). The single intron in the 5’ UTR of the *Penelope* transcript was also shown to reside within the promoter region and to play a role in post-transcriptional regulation. More recently, the 3’ region of PLEs (and therefore also the pLTR) have been shown to contain self-cleaving RNA motifs known as type I hammerhead ribozymes (HHRs) (Cervera and De la Peña 2014). The function of the HHRs is currently unknown.

As with other TE orders, PLEs can be divided into several major clades or superfamilies. The two major EN+ groups are *Neptune* and *Penelope/Poseidon*, which are both found across invertebrate and vertebrate animals and form ancient deep-branching lineages in RT phylogenies (Arkhipova 2006). The two groups also differ in the linker region between the RT thumb and GIY-YIG EN, with *Neptune* elements encoding a 40-50 amino acid extension including a conserved $CX_{2-5}CxxC$ zinc finger motif, which is absent in *Penelope/Poseidon*. A third group present in nematodes and planarian flatworms, *Nematis*, is more closely related to *Penelope/Poseidon* in RT phylogenies but encodes a shorter ~20 amino acid extension including the $CX_{2-5}CxxC$ zinc finger present in *Neptune*. As with major LINE and LTR groups, both *Neptune* and *Penelope/Poseidon* can be further split into well-supported subclades, for example the *Perere* subclade of *Penelope/Poseidon* that is found in flatworms and bdelloid rotifers (Arkhipova et al. 2013). Two additional superfamilies, *Athena* and *Coprina*, are EN– and are generally limited to insertion at the 3’ overhang of telomeric repeats (Gladyshev and Arkhipova 2007). *Athena* elements are restricted to bdelloid rotifers, while the distribution of *Coprina* spans fungi, Archaeplastida (including land plants) and heterokonts, extending the known range of PLEs to at least four eukaryotic kingdoms. *Athena* was also recently shown to include a sub-group of giant retrotransposons called *Terminons*, which can be larger than 40 kb and include multiple ORFs in addition to the PLE RT (Arkhipova et al. 2017). The phylogenetic distribution of PLEs is largely consistent with vertical transmission, with the notable exception of the *Penelope/Poseidon* subclade *Dryad*, that appears to have been horizontally transferred from arthropods to conifers (Lin et al. 2016). *Dryads* are the only described EN+ PLEs outside of animals, although EN+ PLEs of otherwise unknown classification were recently reported from the genome of the streptophyte alga *Chara braunii* (Nishiyama et al. 2018).

In this chapter, I present an exhaustively curated library of TEs for *C. reinhardtii*, which substantially improves existing consensus sequences and incorporates ~140 new TEs. I summarise the TE landscape in *C. reinhardtii*, revealing that the species contains a remarkable diversity of active or recently active TEs. Based on newly described elements in the *C. reinhardtii* library, I describe a novel major clade of PLEs that encode N-terminal GIY-YIG ENs and are distributed across both the green lineage and animals. The evolution of PLEs is discussed.

5.4 Results

5.4.1 An exhaustively curated *Chlamydomonas reinhardtii* transposable element library

To update the existing *C. reinhardtii* TE library, data from several different sources were used. Except for a small number of TEs that were present in only one copy, consensus sequences were manually curated from the alignment of multiple copies of the same TE family or subfamily, following the approach detailed in 1.4.3. In many cases, further support for precise TE termini and target site duplications (TSDs) was obtained by identifying polymorphisms between the reference assembly and a *de novo* assembly of the North Carolina field isolate CC-2931 assembled from Pacific Biosciences (PacBio) long reads (5.6.4). Although not quantified, TE polymorphisms were qualitatively highly abundant between the two assemblies, aiding the annotation of many families. Finally, Iso-Seq data (i.e. PacBio-sequenced cDNAs) were used to curate gene models where relevant, enabling protein sequences to be determined and analysed for a number of TEs that otherwise proved challenging to classify.

As an initial step, all 119 *C. reinhardtii* subfamily consensus sequences (114 families) available from Repbase were re-analysed and new consensus sequences were produced. As well as making minor corrections to existing consensus sequences, I was able to extend and often complete sequences for many models that were previously 5’ truncated or partially annotated. I also extended the classification of 11 sequences to the superfamily level and entirely reclassified 13 sequences (e.g. *TOC1* was erroneously annotated as a DNA transposon and reclassified as a DIRS retrotransposon). An additional nine consensus sequences were produced for TEs described in the literature but absent from Repbase, including the experimentally characterised elements *TOC2* (Day 1995), *REM1* (Perez-Alegre et al. 2005), *MRC1* and *Bill* (Kim et al. 2006). I curated 141 new models, which were largely based on the exhaustive curation of repeat models generated by RepeatModeler (Smit and Hubley 2008-2015). A small number of low-copy number TEs were identified by curating regions displaying presence/absence polymorphisms between the reference assembly and CC-2931, several of which harboured genes in the v5.6 annotation that encoded proteins with putative TE domains. The final updated library contained 269 consensus sequences (224 families), more than doubling the size of the Repbase library both in terms of the number of sequences and the cumulative length of sequences, which increased from 0.48 Mb to 1.09 Mb (Table 1). More than 90% of sequences were classified to at least the order level, with the majority of the unclassified elements belonging to the as of yet uncharacterised "*TE2"* group (e.g. Kapitonov and Jurka (2004b)). When applied to the CC-1690 assembly (O'Donnell et al. 2020), the updated library resulted in an approximately 50% increase in genome-wide TE sequence, increasing from 8.44 Mb to 12.50 Mb (~11.3% of the genome). When combined with other curated sequences from the Volvocales, 12.73 Mb of the genome was identified as TE sequence. Annotation notes for the full updated library are provided in Dataset S1.

Table 1. Comparison between Repbase and updated transposable element libraries.

	Repbase	**Update**
Orders	6	8
Superfamilies	13	16
Families	114	224
Subfamilies	119	269
Classified to at least order (%)	93.28	90.33
Complete sequences (%)	79.83	82.16
Autonomous (%)	46.22	50.93
Cumulative length of sequences (Mb)	0.48	1.09
Total TE sequence in the CC-1690 assembly (Mb)	8.44	12.50

Beyond the number of new TE families discovered, two results derived from the new library are particularly noteworthy. First, *C. reinhardtii* contains an astounding diversity of TEs at the order and superfamily level. The Repbase library includes sequences from six orders (LINEs, LTRs, DIRS, SINEs, DNA transposons and Helitrons), while the updated library extends this to eight with the addition of PLEs and Cryptons (Table 1, Figure 1). Concordantly, the number of superfamilies increased from 13 to 16, with the discovery of *Kyakuja* and *Zisupton* DNA transposons (which I collectively considered part of the superfamily *Kyakuja-Dileera-Zisupton*, or *KDZ*) and the addition of the PLEs and Cryptons. As described in 5.4.2, the PLEs present in *C. reinhardtii* represent a novel superfamily, while the Cryptons were classified to the *CryptonF* superfamily. As a point of comparison, the *Arabidopsis thaliana* TE library includes sequences from five orders and 12 superfamilies, and both the *Drosophila melanogaster* and *Caenorhabditis elegans* libraries include sequences from five orders and 15 superfamilies. Second, TEs in *C. reinhardtii* appear to either be active or very recently active (Figure 1). Approximately 80% of TE sequences identified in the CC-1690 genome were less than 5% divergent from their consensus sequence, which putatively represents the active element for a given family. Indeed, all 16 superfamilies included at least one family displaying an insertion polymorphism relative to CC-2931, transcription supported by Iso-Seq, or both.

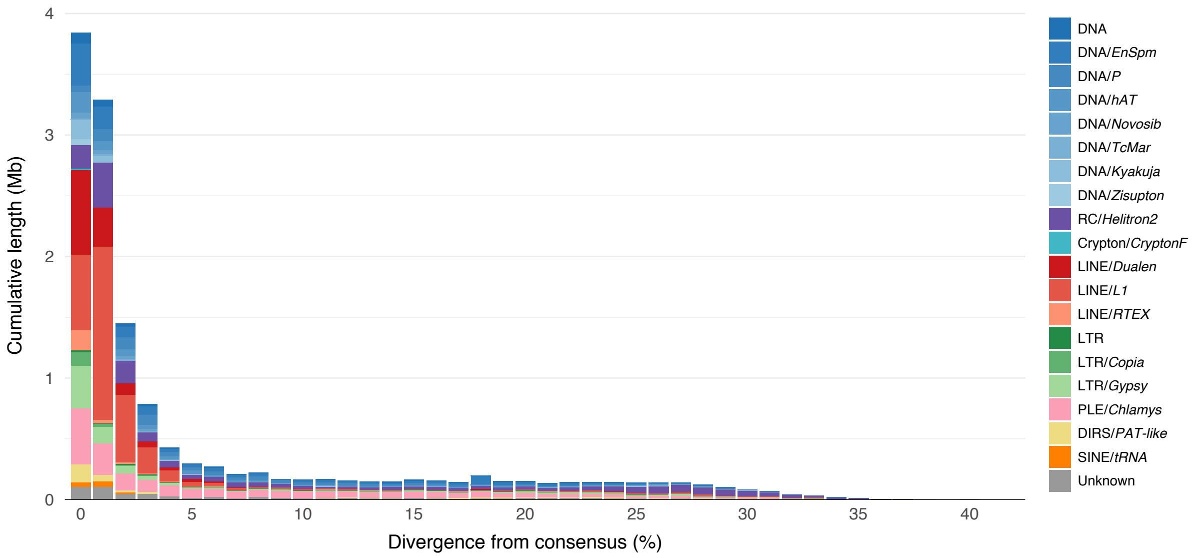


Figure 1. Transposable element landscape for *C. reinhardtii* CC-1690.
Cumulative length of TE sequence plotted against divergence from consensus sequences. TE sequence is coloured by order and superfamily.

At the level of individual families, several results can be highlighted. *Tcr1* was originally identified as a cut-and-paste element producing 8 bp TSDs (Schnell and Lefebvre 1993; Ferris et al. 1996), and Repbase includes a 164 bp fragment of *Tcr1* that is erroneously annotated as an LTR. Although they were unable to sequence the entire *Tcr1* element, Kim et al. (2006) sequenced the complete 314 bp terminal inverted repeats (TIRs), confirming its classification as a DNA transposon. I curated a complete 9,389 bp consensus sequence for *Tcr1*, and using Iso-Seq data I curated two transcribed genes, one of which encodes a transposase from the *Kyakuja* clade. I identified two further autonomous *Kyakuja* families and three autonomous *Zisupton* families, as well as four nonautonomous *KDZ* families. The *Zisupton* elements were particularly remarkable for their length, with *Zisupton-2_cRei* spanning 27,225 bp with TIRs of 1,223 bp, placing it amongst the largest DNA transposons (Arkhipova and Yushenova 2019). Given that *KDZ* was only described in the last decade (Böhne et al. 2012; Iyer et al. 2014), it retrospectively appears that *Tcr1* was likely the first member of this superfamily to be identified. A second DNA transposon, *Tcr3* (Wang et al. 1998), was classified as a *Mariner*-like nonautonomous element in Repbase based on its 2 bp TSDs and lack of obvious ORFs. Once again using Iso-Seq data, I identified three transcribed genes in *Tcr3*, one of which encoded a transposase that resulted in the reclassification of *Tcr3* as an autonomous *EnSpm* element (see Figure 1, Chapter 1).

Considering retrotransposons, the most notable improvements were to LINE elements of the *RTEX* clade (five new autonomous families) and the DIRS elements (six new autonomous families and five new nonautonomous families). All of the DIRS elements were preliminarily assigned to the *PAT-like* superfamily based on the structure of their split repeats, according to the classification system of Poulter and Butler (2015) and Ribeiro et al. (2019). However, it is unclear whether *PAT*-like elements are monophyletic in either RT or RNaseH (RH) phylogenies (Ribeiro et al. 2019) and the tyrosine recombinase (YR) domains of the *C. reinhardtii* DIRS elements form at least two distinct clades (Aaron Vogan, pers. comm.), suggesting that the *PAT-like* assemblage likely requires future revision. All Helitron elements were classified to the *Helitron2* group. These include the three nonautonomous elements encoding Fanzor1 proteins (Bao and Jurka 2013), which are RuvC-like nucleases that likely facilitate transposition (Kojima 2020). Two large autonomous elements, *Helitron-2_cRei* (24,045 bp) and *Helitron-3_cRei* (18,292 bp), encode TET-JBP proteins as discussed in 4.4.8. Large Helitrons encoding secondary genes and gene fragments are not unusual (Kapitonov and Jurka 2007; Castanera et al. 2014), although TET-JBP proteins have so far only been reported to be associated with DNA transposons (Iyer et al. 2014). The *C. reinhardtii* Cryptons were described from just two autonomous families present in the reference genome, both of which were single-copy. However, the individual copies of each element were polymorphic relative to CC-2931, and Iso-Seq data derived from *CryptonF-1_cRei* enabled the identification of a gene encoding a YR domain. Comparison of the YR domain to known Cryptons and the identification of an additional GCR1_C DNA-binding domain placed these elements in the *CryptonF* superfamily, which has been previously described from fungi and oomycetes (Kojima and Jurka 2011).

5.4.2 *Chlamys*: a major new clade of *Penelope*-like elements

The most surprising result from the updated library was the discovery of ten autonomous retrotransposon families encoding RT domains with homology to PLEs. Given the novelty of this result, I performed targeted annotation of similar TEs in the genomes of *Chlamydomonas incerta*, *Chlamydomonas schloesseri* and *Edaphochlamys debaryana* (3.6.5), curating 29 families across the four species (Dataset S2). All of these families exhibited genome-wide distributions, unlike almost all PLEs found outside of animals, which are EN– and telomere-restricted (Gladyshev and Arkhipova 2007). However, the predicted ORFs did not contain C-terminal GIY-YIG EN domains as found in EN+ PLEs (Arkhipova 2006). Individual insertions of the families were often heavily 5' truncated, although curation of the most complete copies revealed the presence of an N-terminal GIY-YIG EN domain in 13 of the families. The remaining families presumably also encode N-terminal ENs (enabling genome-wide retrotransposition), which are missing due to the absence of full-length copies in the reference genomes. pLTRs were detected in at least one copy for seven of the families, suggesting that these elements are full-length and could potentially be active in the respective reference strains. As presented in detail below, the protein sequences of both the RT and EN domains shared multiple features with those of canonical PLEs, as well as several notable differences. Considered collectively, the presence of RT domains with homology to PLEs and of GIY-YIG ENs, together with the observed pLTRs and 5' truncation, clearly indicates that these new elements represent a novel type of PLE. I name these elements *Chlamys*, following the previous examples of *Coprina* and *Nematis*, which were both named for the taxon in which they were discovered.

The full-length *Chlamys* elements ranged from 7.29 kb to 4.36 kb (not including pLTRs) and specific insertion at $(C)_n$ or $(CA)_n$ microsatellites was observed in several families (Dataset S2). Length variation was largely driven by the presence of additional protein domains or satellite DNA. If present, satellite DNA was generally located downstream of the ORF or for seven families within the ORF, specifically within the DKG domain (see Arkhipova (2006) and below). For example, the families *Chlamys-8_cRei* (6.57 kb, complete) and *Chlamys-2_cSch* (7.20 kb, incomplete) both contained satellite repeats starting at a region corresponding to ~25 amino acids downstream of the conserved DK motif. While the satellite repeat is in-frame in *Chlamys-8_cRei* (which would result in an ~450 amino acid insertion), the *Chlamys-2_cSch* repeat is not. It is possible that these families do not produce functional proteins and are non-autonomous, utilising the proteins produced by families lacking internal satellite repeats for their retrotransposition. However, there is no evidence that these families are more degraded in regions encoding highly conserved amino acids, which would be expected if they had been pseudogenised.

Searching for *Chlamys* elements in further chlorophyte genomes yielded four additional families (three complete, one truncated) in the following species: *Raphidocelis subcapitata*, *Tetradesmus obliquus* and *Desmodesmus armatus* (all class Chlorophyceae, order Sphaeropleales), and *Botryococcus braunii* (class Trebouxiophyceae). The Sphaeropleales are estimated to have diverged from the lineage leading to *Chlamydomonas* ~600 million years ago, while the Trebouxiophyceae likely diverged 700-1,000 million years ago (Del Cortona et al. 2020). *Chlamys-1_rSub* contains a satellite array within EN, between the core GIY-YIG motif and the conserved E and N amino acids (see Dunin-Horkawicz et al. (2006) and below). ORFs for the other three elements were uninterrupted. At 8.18 kb, *Chlamys-1_bBra* is the longest element annotated in this study. This family contains a putative 362 amino acid ORF upstream of the main EN-RT ORF with no known domains or blastp hits, and an ~2 kb region at the 3' end consisting of tandem repeats. Extending the search to the streptophyte lineage (land plants and their green algal relatives) revealed one family in each of the unicellular algae *Mesostigma viride* and *Chlorokybus atmophyticus*. The complete consensus sequences for these families (*Chlamys-1_meVir* 4.55 kb, *Chlamys-1_cAtm* 3.76 kb) encoded uninterrupted ORFs and did not appear to be outstanding in any respect relative to the families from Chlorophyta. Interestingly, HHR motifs were not detected in any of the *Chlamys* families. All *Chlamys* elements were recovered as a strongly supported clade in an RT phylogenetic analysis, with a branching order consistent with vertical transmission (Figure 2). As with the evolutionary relationships of their hosts, the streptophyte elements formed a sister group to the chlorophyte elements, the single element from Trebouxiophyceae (*Chlamys-1_bBra*) was an outgroup to the elements from Chlorophyceae (Volvocales + Sphaeropleales), and the Sphaeropleales elements were outgroups to the Volvocales elements from *Chlamydomonas* and *Edaphochlamys*. Thus, *Chlamys* represents a new superfamily of

PLEs that was likely present in the common ancestor of Viridiplantae over a billion years ago.

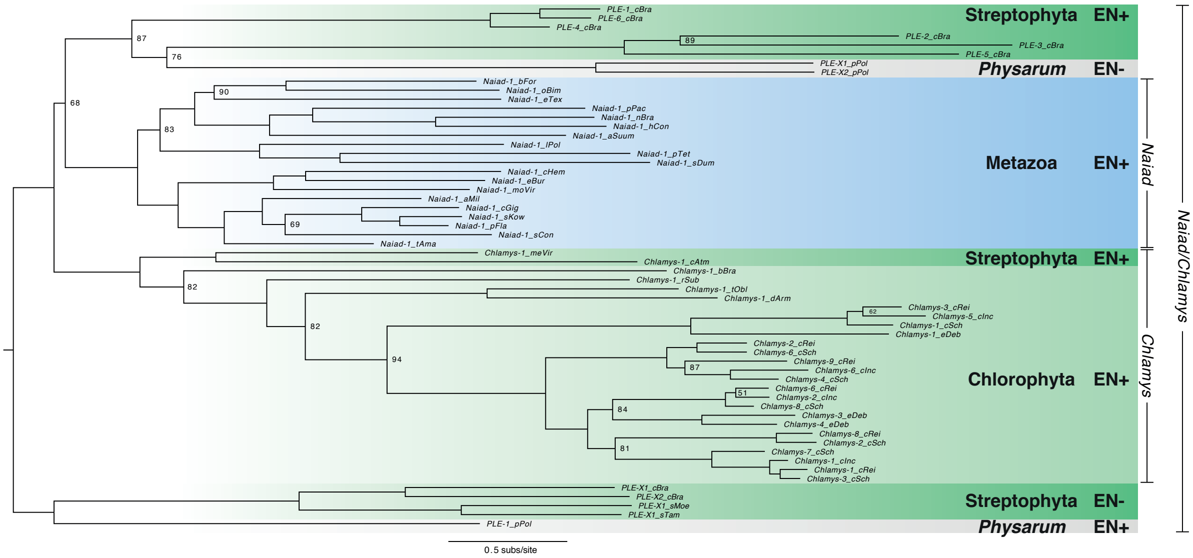


Figure 2. Reverse transcriptase phylogeny of the *Naiad/Chlamys* clade of N-terminal EN+ PLEs. Phylogeny of *Chlamys, Naiad* and *Chlamys*-like elements based on alignment of RT protein sequences from the CxC domain to the RT thumb. The phylogeny was rooted on *Athena* core RT/thumb sequences (not shown). Bootstrap values <95% are shown.

The predicted peptides for all *Chlamys* families shared several features, namely the N-terminal GIY-YIG EN, a variable length linker region between EN and RT, the N-terminal RT region, core RT, and the RT thumb/C-terminal region (see *Chlamys-1_meVir* example, Figure 3). The GIY-YIG domain of *Chlamys* elements differed markedly from previously described PLEs, with both the CX$_{2}$-$_{5}$CxxC zinc-finger motif located upstream of the GIY-YIG motif and the CCHH zinc-finger motif centred on the GIY-YIG motif absent (Figures 4, S1). While the CX$_{2}$-$_{5}$CxxC motif is absent from the *Penelope/Poseidon* group, the CCHH motif is a defining feature of GIY-YIG ENs found in canonical PLEs (Arkhipova 2006). The linker region immediately following the GIY-YIG EN was rich in R and K amino acids. The sequence immediately preceding the N-terminal RT region contained several well-conserved residues and was characterised by a perfectly conserved domain with a CxC motif, followed by a conserved histidine 12-17 amino acids downstream. This domain is absent from all described PLEs and appears to be a distinguishing feature of *Chlamys* (and other N-terminal EN+ PLEs, 5.4.4, 5.4.5). Immediately following the novel N-terminal domain, *Chlamys* contained the RT N-terminal regions corresponding to the previously described N3, N2 and N1 domains (Arkhipova 2006). All *Chlamys* elements contained a perfectly conserved DK motif upstream of the core RT, which is present in all PLEs and is within the larger region described as the DKG domain (Arkhipova 2006). None of the peptides included a G immediately after the DK, although this residue is not perfectly conserved in described PLEs and is commonly substituted in *Athena* and *Coprina*. Most of the core RT domains and the RT thumb were unremarkable when compared to those of PLEs and TERTs, with the exception of a region between RT3(A) and RT4(B) known as the insertion in the fingers domain (IFD) (Lingner et al. 1997; Lue et al. 2003). The IFD can be split into three domains, IFDa and IFDc are present in the RTs of PLEs, TERTs and Group II introns, while IFDb (also known as the TRAP domain) is an ~90 amino acid linker between IFDa and IFDc that is unique to TERTs (Jiang et al. 2018). Remarkably, *Chlamys* elements contained an ~40 amino acids insertion between IFDa and IFDc (Figure S2), potentially strengthening the evolutionary link between PLEs and TERTs.

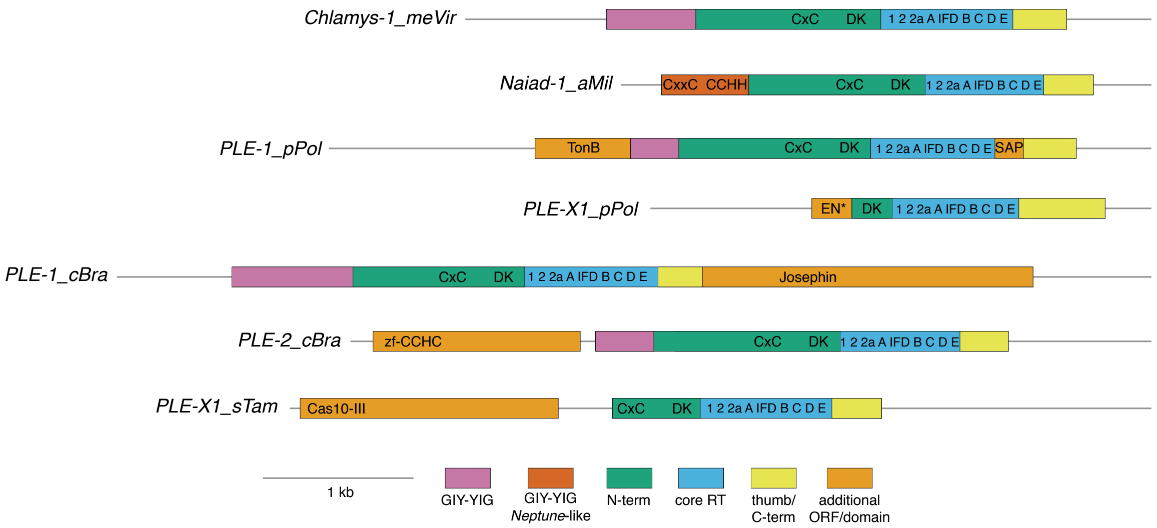


Figure 3. The diversity of *Chlamys, Naiad* and *Chlamys*-like elements.
Schematics of *Naiad/Chlamys* elements. All examples are from complete consensus sequences and pLTRs are not shown. All domains are introduced in the main text except EN* in *PLE-X1_pPol*, which is a possible degraded remnant of the GIY-YIG motif.

Several families encoded additional domains beyond the minimal configuration described above. Four families (*Chlamys-1_cRei*, *Chlamys-1_cInc*, *Chlamys-3_cSch* and *Chlamys-7_cSch*) shared an ~100 amino acid insertion containing no known domains between the RT N-terminal region and RT1. Three families (*Chlamys-3_cRei*, *Chlamys-5_cInc* and *Chlamys-1_cSch)* contained a plant homeodomain (PHD) finger inserted within RT2a, and an additional N-terminal extension upstream of the GIY-YIG EN with no known domains. Two families (*Chlamys-8_cRei* and *Chlamys-2_cSch*) contained a PHD finger domain within the EN, between the GIY-YIG domain and the conserved E and N. The PHD finger contains a conserved zinc-finger C4-H-C3 motif, and recognises the histone H3 in a sequence and modification dependent manner (Sanchez and Zhou 2011). PHD fingers have been reported from a variety of TEs (e.g. in certain *CR1* LINE elements (Kapitonov and Jurka 2003) and *Rehavkus* DNA transposons (Dupeyron et al. 2019)), where they may play a role in chromatin restructuring. A PHD finger domain has previously been reported in the *C. reinhardtii* Gypsy LTR element *REM1* (Perez-Alegre et al. 2005), and I also identified PHD fingers in at least two additional Gypsy families (including *Gypsy-7a_cRei*, 4.4.6), the DIRS element *PAT-6_cRei* and the DNA transposon *Kyakuja-2_cRei* (Dataset S1). The PHD finger may therefore be considered a general accessory domain in *C. reinhardtii* TEs, where it may play a role in targeting insertions to genomic regions unlikely to disrupt genes.

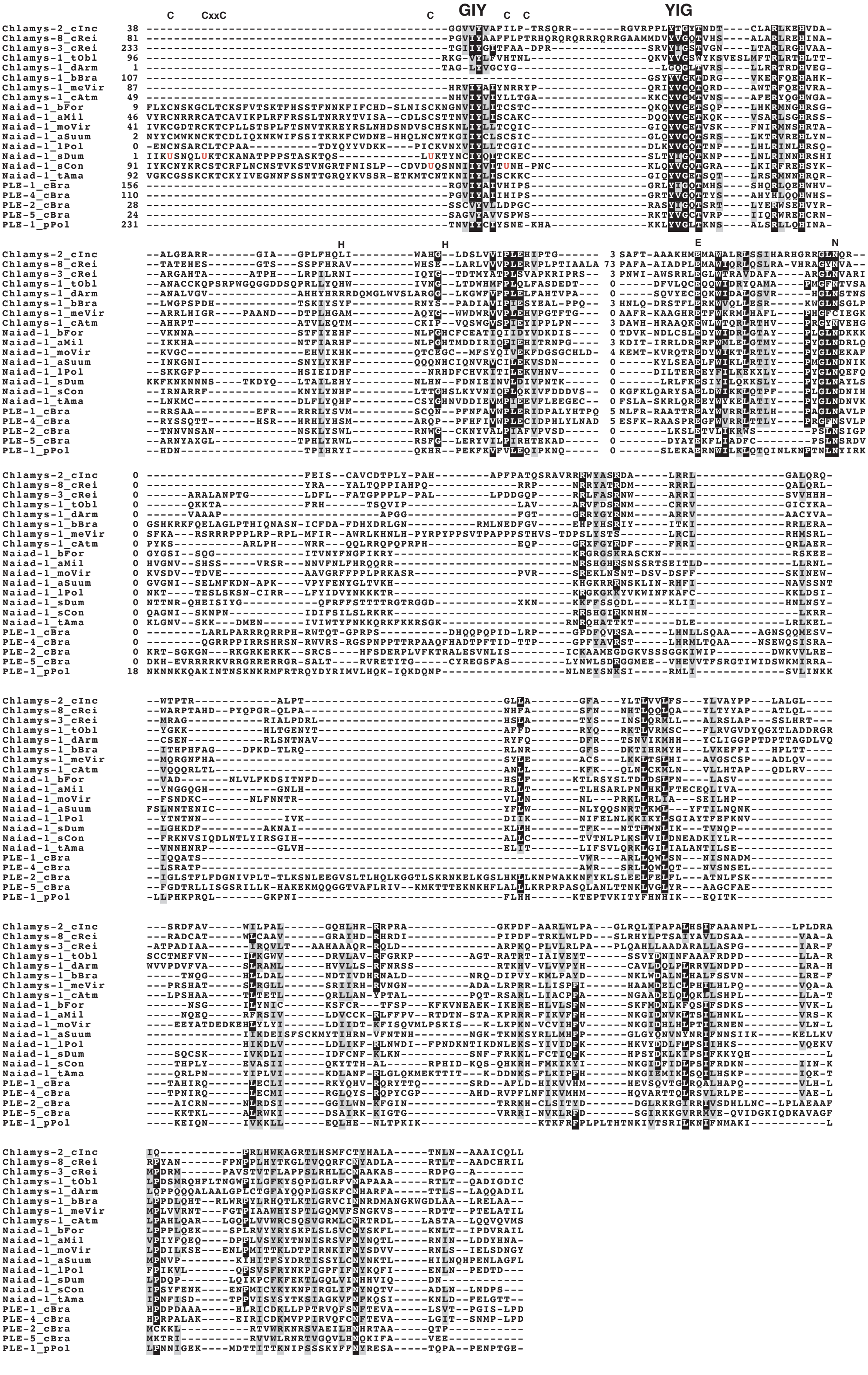


Figure 4 (continues next page).

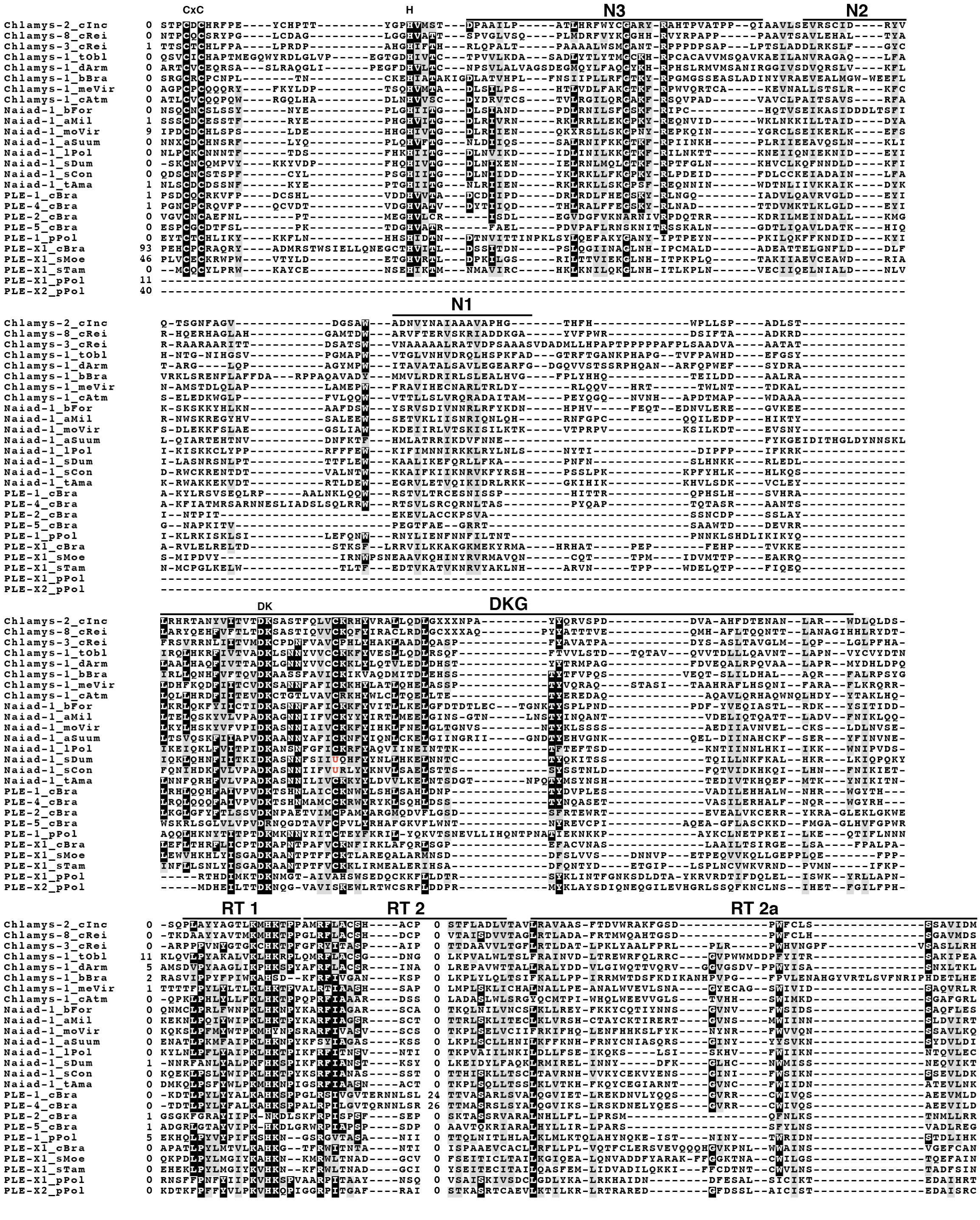


Figure 4 (continues next page).

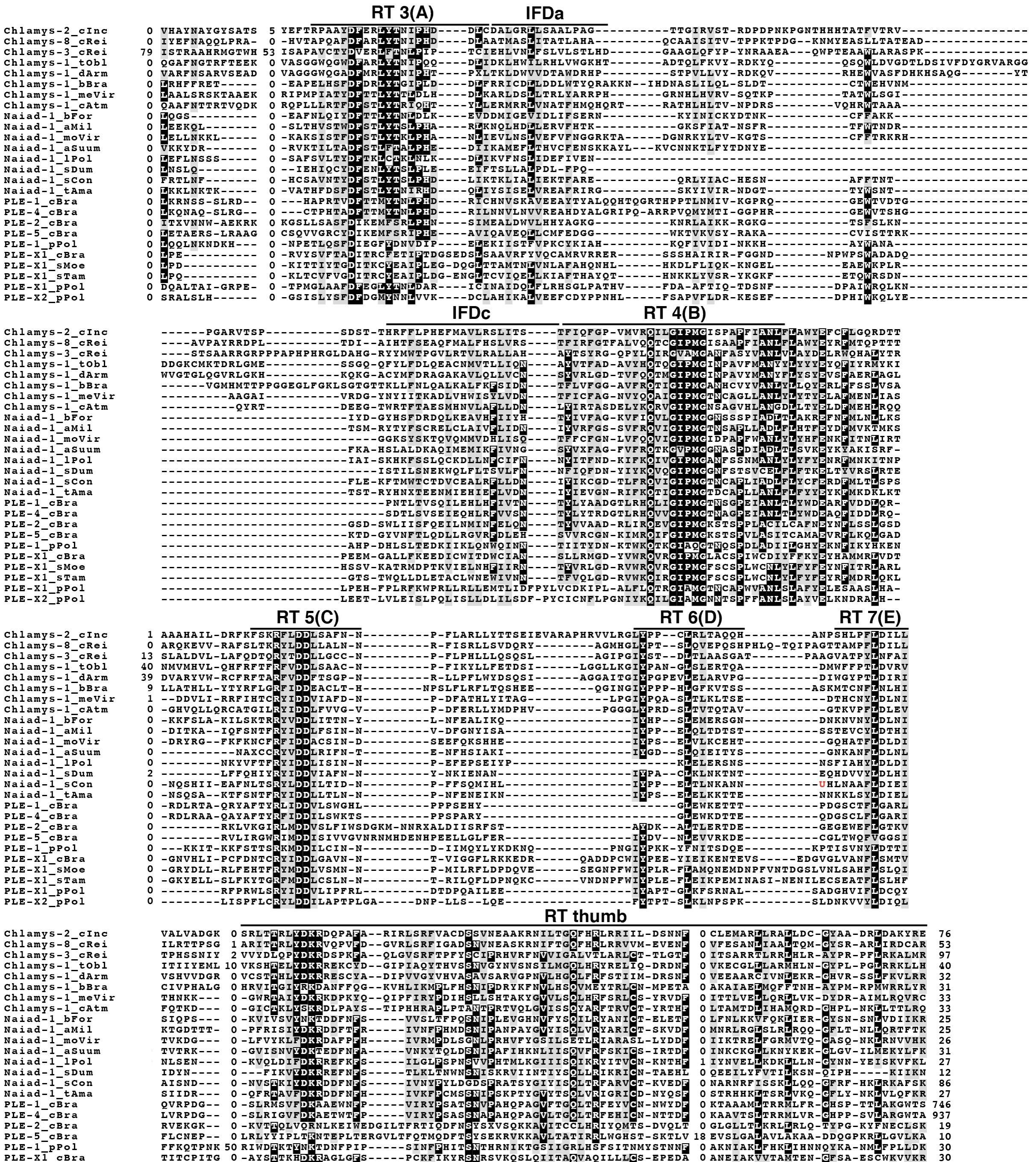


Figure 4. Alignment of *Chlamys, Naiad* and *Chlamys*-like protein sequences.
Domains are highlighted according to Arkhipova (2006), with the exception of IFDa and IFDc (Jiang et al. 2018). Selenocysteines are highlighted in red text. The conserved C and H residues in the $CX_{2-5}CxxC$ and CCHH motifs found in *Naiads* are marked in the EN domain. Note that EN– elements are not included on the first page. Alignment was produced with PROMALS3D (Pei and Grishin 2014).

5.4.3 Functional characterisation of a *Chlamys* element

Based on the functional genomics resources available for *C. reinhardtii,* I further focussed on the autonomous elements identified in the species. As with most *C. reinhardtii* TEs (Figure 1), *Chlamys* copies exhibited minimal divergence from their respective consensus sequences, suggesting recent activity (Figure S3A). This was confirmed by comparison to CC-2931, with polymorphic insertions observed for copies of all ten families. The autonomous *Chlamys* elements cumulatively spanned 1.07 Mb of the CC-1690 assembly. The 6.89 kb family *Chlamys-3_cRei* was the most abundant, being present in 276 copies and spanning 0.19 Mb. However, only 43 copies were >1 kb in length, demonstrating the extent of 5' truncation (Figure S3B). Only two families, *Chlamys-3_cRei* and *Chlamys-8_cRei*, contained complete copies with pLTRs.

I also curated models for eight families of putatively nonautonomous *Chlamys* elements, five of which contained pLTRs, which collectively spanned 0.77 Mb of the genome. These elements most commonly displayed sequence similarity to autonomous families at their 3' end and exhibited similar insertion biases at $(C)_n$ or $(CA)_n$ microsatellites. They include the experimentally characterised *MRC1* (Kim et al. 2006), which often inserts in multiple head-to-tail tandem copies (4.4.6) and may generally be the most active TE across laboratory strains (Gallaher et al. 2015; Neupert et al. 2020). *MRC1* exhibits 96.4% sequence similarity over 137 bp at its 5' end to the 5' of *Chlamys-8_cRei*, making it the only nonautonomous element exhibiting 5' homology. As with *Chlamys-8_cRei*, *MRC1* does not appear to have any target site preference. Since *Chlamys-8_cRei* is one of only two families with complete copies in the reference genome, it is a strong candidate to be the autonomous partner of *MRC1*, although I did not observe transcription from the complete copy. Interestingly, *Chlamys-8_cRei* is among the families containing internal satellite DNA (5.4.2), suggesting that at least some of these families may produce functional proteins.

For the one other family with a complete insertion in the reference genome, *Chlamys-3_cRei*, I observed transcription in the Iso-Seq data (Figure 5A). Unfortunately, this copy contains a 2.84 kb deletion relative to the consensus. Fortunately, the deletion is entirely within the EN-RT ORF, and so the copy presumably retains a functional promoter, TSS and terminator. Remarkably, the transcription and derived gene model of *Chlamys-3_cRei* shares many features with that of *Penelope* from *D. virilis*, the only functionally characterised PLE. As introduced in 5.3, the *Penelope* TSS and internal promoter are located within the pLTR, and while the RT-EN ORF initiates within the complete *Penelope* body (i.e. downstream of the pLTR), the 5' UTR within the pLTR contains a 75 bp intron that overlaps the internal promoter region (Arkhipova et al. 2003; Schostak et al. 2008). The *Chlamys-3_cRei* TSS is also located within the pLTR, while a 398 bp intron in the 5' UTR with valid GT/AG donor and acceptor sites spans the boundary between the pLTR and the downstream main body (Figure 5A, B). A peak in mapped H3K4me3 ChIP-seq data indicates the presence of an internal promoter (Ngan et al. 2015), which largely overlaps the intron. The longest ORF initiates 41 bp downstream of the intron within the main body of the element. Additionally, three Iso-Seq reads support an alternative transcript with a 751 bp intron, which shares the donor site of the 398 bp intron but has an alternative AG acceptor site. This transcript initiates at a downstream in-frame start codon and results in a peptide 293 amino acids shorter than the more abundant transcript. As the predicted *Chlamys-3_cRei* peptide contains an N-terminal extension, both ORFs encode complete EN and RT domains and are potentially functional (Figure 5B). The similarities between *Penelope* and *Chlamys-3_cRei* potentially indicates an ancient and deeply conserved organisation and mechanism shared by canonical PLEs and *Chlamys* elements.

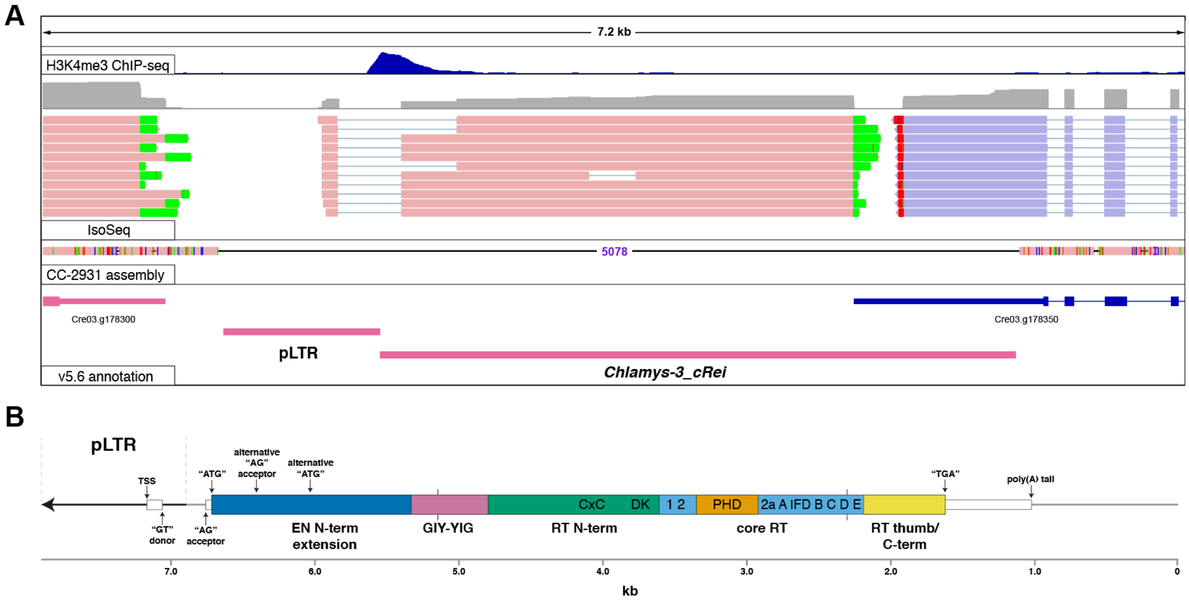


Figure 5. Functional characterisation of *Chlamys-3_cRei*.
(A) IGV browser view of a *Chlamys-3_cRei* copy that is polymorphic relative to the CC-2931 assembly. Green and red mismatched bases on Iso-Seq reads represent poly(A) tails of transcripts. Iso-Seq and ChIP-seq data were obtained from Gallaher et al. (2021).
(B) Schematic of the structural organisation of *Chlamys-3_cRei*. Note that this represents the full-length element and the transcribed copy above contains a 2.84 kb internal deletion, the boundaries of which are shown by the dashed black lines.

5.4.4 *Naiad*: a metazoan clade of *Penelope*-like elements with N-terminal endonuclease

Beyond green algae, I identified many putative TEs with homology to *Chlamys* elements in animal species. Curation of these elements revealed that, like *Chlamys*, they encode ORFs containing an N-terminal GIY-YIG EN and RT, and display all of the expected characteristics of PLEs. I produced consensus sequences for a single family (usually the most abundant) from each of 19 species. Of these 19 families, eight consensus sequences were complete (i.e. at least one copy had a pLTR) and nine were near complete (i.e. the predicted ORF contained an N-terminal GIY-YIG EN), leaving only two families that were based on homology of the RT domain alone. Curation effort was focused to maximise taxonomic diversity, which resulted in families annotated from 7 phyla and 14 classes (Dataset S3). Elements were annotated from Ctenophora (the cigar comb jelly *Beroe forskalii*), Cnidaria (the hydrozoan *Clytia hemisphaerica,* stony coral *Acropora millepora* and box jellyfish

Morbakka virulenta), Nematoda (*Pristionchus pacificus, Nippostrongylus brasiliensis, Haemonchus contortus* and *Ascaris suum*), Arthropoda (the Atlantic horseshoe crab *Limulus polyphemus,* American house spider *Parasteatoda tepidariorum,* African social spider *Stegodyphus dumicola* and Texas clam shrimp *Eulimnadia texana*), Mollusca (Pacific Oyster *Crassostrea gigas*, Chinese razor clam *Sinonovacula constricta* and the California two-spot octopus *Octopus bimaculoides*), Hemichodata (the acorn worms *Saccoglossus kowalevskii* and *Ptychodera flava*) and Chordata (the inshore hagfish *Eptatretus burgeri* and prehistoric monster fish *Thalassophryne amazonica*). The elements from *C. gigas* and *E. texana* were based on models found in Repbase (*Penelope-2_CGi* and *Penelope-1_EuTe*), although both Repbase models were 5' truncated and did not include EN domains, making them difficult to distinguish from canonical EN– PLEs without phylogenetic analysis. I did not find any elements in many of the most well-studied animal taxa (e.g. tetrapods, insects and *Caenorhabditis*), likely explaining why these elements have been missed in previous annotation efforts. Aside from the two fish species above, I was unable to find any elements in other vertebrates. The elements annotated from *E. burgderi* and *T. amazonica* are present in multiple copies and are found on well-assembled contigs, suggesting that they do not represent sequencing contamination.

The families from animal genomes were generally shorter than those from green algae, with complete elements ranging from 3.38 kb to 4.37 kb. No satellite arrays were present and no protein domains or insertions aside from the minimal organisation were detected. Five families had clear $(TA)_n$ insertion targets (Dataset S3). As with *Chlamys*, no HHR motifs were detected in any of the families. Interestingly, several complete families lacked a start codon. Assuming the transcription start site of these elements is present in the pLTR and the expectation of an intron, as found in *Penelope* and *Chlamys-3_cRei*, it is possible that the start codon is located in the pLTR and/or in a different frame. Based on the RT phylogeny, all of the new metazoan elements formed a strongly supported clade (Figure 2). Several elements formed subclades congruent with their host taxa, for example the four elements from nematodes and three elements from chelicerates. Although incongruence was observed for several other taxa (e.g. the two elements from fish), this does not necessarily imply horizontal transfer. As the families were essentially curated at random, it may simply be that any two families had already diverged in ancient animal evolution, long before the most recent common ancestor of the clade of interest (e.g. vertebrates for the two fish families). Following personal communication with Irina Arkhipova, this new clade is named *Naiad* for the water nymphs from Greek mythology, following the aquatic habitat of many of their hosts. The 3,517 bp family *Naiad-1_aMil* is representative of the clade (Figure 3).

The most remarkable feature of the *Naiad* peptides was that the GIY-YIG EN closely resembled those present in C-terminal EN+ PLEs, particularly *Neptune*. The putative peptides from all complete elements had both the $CX_{2-5}CxxC$ linker motif upstream of the GIY-YIG domain and the CCHH zinc-finger motif within the GIY-YIG domain (Figures 4, S1). The novel CxC domain upstream of the RT N-terminal region was present as in *Chlamys*. All peptides contained the DK motif. The insertion between IFDa and IFDc was also observed, although at ~20 amino acids it was approximately half the length of the insertion found in

Chlamys. The three elements from chelicerates (*Naiad-1_lPol, Naiad-1_pTet* and *Naiad-1_sDum*) appear to have entirely lost this insertion, and essentially resembled canonical PLEs in this respect (Figure 4). Although *Chlamys* and *Naiad* are phylogenetically related (Figure 2, 5.4.7), I propose individual classifications based primarily on the distinguishing features of the GIY-YIG EN, and additionally the length variation observed for the IFD insertion. As it stands these clades are obviously also defined based on taxonomic restriction to animals or green algae, although this may be revised based on future annotation.

Finally, the ORFs of two families, *Naiad-1_sCon* from the Chinese razor clam and *Naiad-1_sDum* from the African social spider, each contained four in-frame UGA codons. Collectively, seven of the eight UGA codons corresponded to highly conserved cysteines in the peptide sequence (Figure 4). Three of the four conserved cysteines in the $CX_{2-5}CxxC$ zinc-finger motif corresponded to UGA codons in *Naiad-1_sDum*, and one of four in *Naiad-1_sCon*. In *Naiad-1_sCon*, the first cysteine in the CCHH motif corresponded to UGA, as did the conserved cysteine in the DK motif in both families. One possible explanation is stop codon read-through. UGA stop codons are the 'leakiest' in many animal species, where either arginine, cysteine, serine or tryptophan have been shown to be incorporated (Jungreis et al. 2011). Alternatively, the UGA codons in these families may be encoding selenocysteine. The incorporation of selenocysteine requires the recoding of the UGA codon from stop, which in eukaryotes is achieved by recognition of the selenocysteine insertion sequence (SECIS) located in the 3' UTR of the selenoprotein mRNAs (Low and Berry 1996). Using SECISearch3 (Mariotti et al. 2013), I identified "grade A" (i.e. the highest confidence) SECIS elements in the consensus sequences of both elements (Figure S4). In *Naiad-1_sCon* the SECIS starts 21 bp downstream of the putative stop codon, while in *Naiad-1_sDum* it starts just 2 bp from the stop codon, presumably placing the SECIS in the 3' UTR for both elements. It therefore appears that the EN-RT product of both families is a selenoprotein. Furthermore, based on the phylogenetic relationship of the families (Figure 2), it is expected that the evolutionary transition to selenoproteins has occurred independently in each lineage.

5.4.5 *Chlamys*-like elements, with and without endonuclease

I also identified a small number of families that differed from *Chlamys* and *Naiad*, but were more closely related to them than to previously described PLEs. These *Chlamys*-like elements often contained additional domains or secondary ORFs, and included both N-terminal EN+ and EN– families (Dataset S4). Three families were identified in the slime mold *Physarum polycephalum*. One of these elements, *PLE-1_pPol*, encodes a 1,197 amino acid peptide containing an N-terminal GIY-YIG EN (with neither the $CX_{2-5}CxxC$ or CCHH motifs), the *Naiad/Chlamys* CxC domain and DK motif, and a core RT with an ~20 amino acid insertion between IFDa and IFDb (Figure 4). The *PLE-1_pPol* peptide contains two additional domains, one matching the TonB superfamily located upstream of the GIY-YIG EN, and a second SAP domain inserted between RT7(E) and the RT thumb (Figure 3). TonB is a bacterial membrane protein involved in transport of siderophores and other molecules (Noinaj et al. 2010), and it is unclear what role this domain plays in *PLE-1_pPol*. SAP (SAF A/B, Acinus and PIAS) is a putative DNA-binding motif (Aravind and Koonin 2000) that has previously been reported in Zisupton DNA transposons (Böhne et al. 2012). It is possible that the SAP domain plays some role in target site specificity, and *PLE-1_pPol* does exhibit insertion at $(CA)_n$ microsatellites (although microsatellite insertions are common in *Chlamys* and *Naiad* elements without SAP domains). The two other *P. physarum* families, *PLE-X1_pPol* and *PLE-X2_pPol*, encoded peptides that were truncated at the N-terminus, with almost the entire RT N-terminal region absent beyond DK (including the CxC domain) and only a degraded possible remnant of the GIY-YIG motif present (Figure 3 & 4). *PLE-X1_pPol* corresponds to the previously described unclassified element *Physarum*, which was identified as a 5' truncated family and was notable for its isolated phylogenetic position and IFD insertion (Gladyshev and Arkhipova 2007). Telomeric repeats were not identified flanking either *PLE-X1_pPol* or *PLE-X2_pPol*, both of which appeared to be associated with AT-rich regions. Despite the presence of a possible EN remnant, I classify these elements as EN– based on their N-terminal truncation.

The remaining families were identified in streptophyte species, although they were distinct from the described *Chlamys* clade. In line with the report of PLEs in the *C. braunii* genome (Nishiyama et al. 2018), I identified eight PLE families in the species. These appeared to fall into three distinct groups. The first, represented by *PLE-1_cBra*, *PLE-4_cBra* and *PLE-6_cBra*, encoded long ORFs of 1,768 - 1,936 amino acids. These included N-terminal GIY-YIG EN (lacking $CX_{2-5}CxxC$ and CCHH motifs), the CxC domain, an IFD insertion of ~30 amino acids (Figure 4), and long C-terminal extensions including a putative Josephin domain (Figure 3). Josephin domains function in de-ubiquitination (Tzvetkov and Breuer 2007) and Josephin-related cysteine protease domains are a feature of *Dualen* LINE elements, where they potentially disrupt protein degradation (Kojima and Fujiwara 2005). The second group, comprising *PLE-2_cBra*, *PLE-3_cBra* and *PLE-5_cBra*, encoded a relatively standard EN-RT ORF (no $CX_{2-5}CxxC$ or CCHH in the GIY-YIG EN) and an additional upstream ORF of 401 - 458 amino acids including a gag-like zinc-knuckle domain (zf-CCHC) in *PLE-2_cBra* (Figure 3) and *PLE-5_cBra*.

The third group included two families from *C. braunii* (*PLE-X1_cBra* and *PLE-X2_cBra*) and one family each from the lycophytes *Selaginella moellendorffii* (*PLE-X1_sMoe*) and *Selaginella tamarschina* (*PLE-X1_sTam*). These families are EN–, although unlike the *P. physarum* EN– families their peptides do include the CxC domain (Figure 4). These families contain an upstream ORF of 486 – 571 amino acids, ~100 amino acids of which aligns consistently among the four elements and includes four conserved cysteines. The upstream ORF of *PLE-X1_sTam* contains an additional putative domain with homology to CRISPR/Cas system-associated protein Cas10 (Figure 3). It is possible that this a spurious hit as this region partially aligns to the upstream ORF of *PLE-X1_sMoe*, which lacks conservation of the amino acids that result in this domain being predicted in *PLE-X1_sTam*. Interestingly, *PLE-X1_sTam* is remarkable in that all copies were found to share precisely the same insertion site within the 28s ribosomal RNA gene (in reverse orientation at 1,128 bp relative to the partial 28s ribosomal RNA gene of *Selaginella stauntoniana*, NCBI accession AJ507613.1). Targeted insertion at specific sites of the 28s rRNA genes is known in Arthropods from *R1* and *R2* LINE elements (Eickbush 2002) and *Pokey* DNA transposons

(Penton and Crease 2004), which is thought to have evolved as ribosomal DNA represents a safe "habitat" for insertion given its tandem duplication and high copy number. The other three families in this EN– group were associated with neither ribosomal DNA nor telomeric sequence.

The phylogenetic relationships of these additional families relative to *Chlamys* and *Naiad* were generally uncertain. The EN+ *PLE-1_pPol* formed a well-supported clade with the EN– families from *C. braunii* and the *Selaginella* species. The remaining families form three clades, consisting of two EN– families from *P. physarum*, the three EN+ *C. braunii* families with Josephin domains, and the three EN+ *C. braunii* families with additional upstream ORFs (Figure 2, Figure 3). Interestingly, HHR motifs were identified in *PLE-2_cBra* and *PLE-X1_cBra*, but in no other elements. As demonstrated in 5.4.7, these families clearly form a major clade with *Chlamys* and *Naiad* to the exception of all other described PLEs. However, I suggest that these poorly represented subclades be referred to simply as *Chlamys*-like until additional elements are described, and their phylogenetic relationships are further elucidated. I use *Naiad/Chlamys* to refer to the entire clade of elements presented in Figure 2.

5.4.6 *Hydra*: an enigmatic clade of metazoan *Penelope*-like elements with C-terminal endonuclease

When comparing the novel PLEs to previously described PLEs (5.4.7), I noticed seven C-terminal EN+ families from Repbase that formed an isolated group and appeared to encode peptides substantially divergent from *Neptune, Nematis* and *Penelope/Poseidon*. Six of these families were annotated from the freshwater polyp *Hydra magnipapillata* and one was from the starlet sea anemone *Nematostella vectensis*. Once again focussing on taxonomic diversity, I used the peptides from these elements to identify and curate six related families, two in the stony coral *A. millepora*, two in the sea cucumber *Apostichopus japonicus*, and one each in the California two-spot octopus *Octopus bimaculoides* and sea louse *Caligus rogercresseyi* (Dataset S5). These elements, which I refer to as *Hydra*, are generally short (<3 kb) and all exhibited $(TA)_n$ microsatellite insertions. HHRs were identified in the 3' end of all families. At the peptide level, they possess the general features of canonical PLEs (i.e. no CxC domain upstream of the RT N-terminal region, DK motif present, no IFD insertion, C-terminal GIY-YIG present), although they are highly divergent throughout and some regions align poorly (e.g. RT4(B), Figure S5). Interestingly, the *Hydra* GIY-YIG EN possessed several unique features. A linker region was present between RT and EN, although the four conserved cysteines were arranged in a different configuration to the $CX_{2\text{-}5}CxxC$ motif present in *Neptune* and *Nematis* (Figures S1, S5). The CCHH motif was also absent, although a potential CxxC conserved motif was found upstream of the conserved E and N amino acids.

5.4.7 An updated phylogeny of *Penelope*-like elements

In the recent phylogenetic analysis performed by Arkhipova et al. (2017), canonical PLEs formed a well-supported clade with TERTs as an outgroup. Within the canonical PLEs there were three major clades, namely *Athena*, *Neptune* + *Coprina*, and *Penelope/Poseidon* + *Nematis*. The phylogenetic relationship of these three clades to each other is uncertain. I performed a phylogenetic analysis of the core RT of TERTs, canonical PLEs, *Hydra* and *Naiad/Chlamys* (Figure 6). I recovered the three major clades of canonical PLEs, although the *Penelope/Poseidon* + *Nematis* clade only received 90% ultrafast bootstrap support. The monophyly of *Neptune* was not recovered, in line with Arkhipova et al. (2017). The EN– superfamily *Coprina* was recovered as a clade with 98% support, grouping within the diversity of *Neptune* elements. Collectively, the canonical PLEs were recovered as a clade with 90% support. Beyond the canonical PLEs, there were three additional major clades, TERTs, *Naiad/Chlamys*, and *Hydra*. If TERTs were assumed to be the root, a clade of all PLEs was strongly supported (100%), with *Naiad/Chlamys* forming a basal clade to all other PLEs (canonical PLEs + *Hydra*). However, *Hydra* was present on a long branch with very low ultrafast bootstrap support for its grouping with canonical PLEs (41%), and its evolutionary relationship with other PLEs is unresolved.

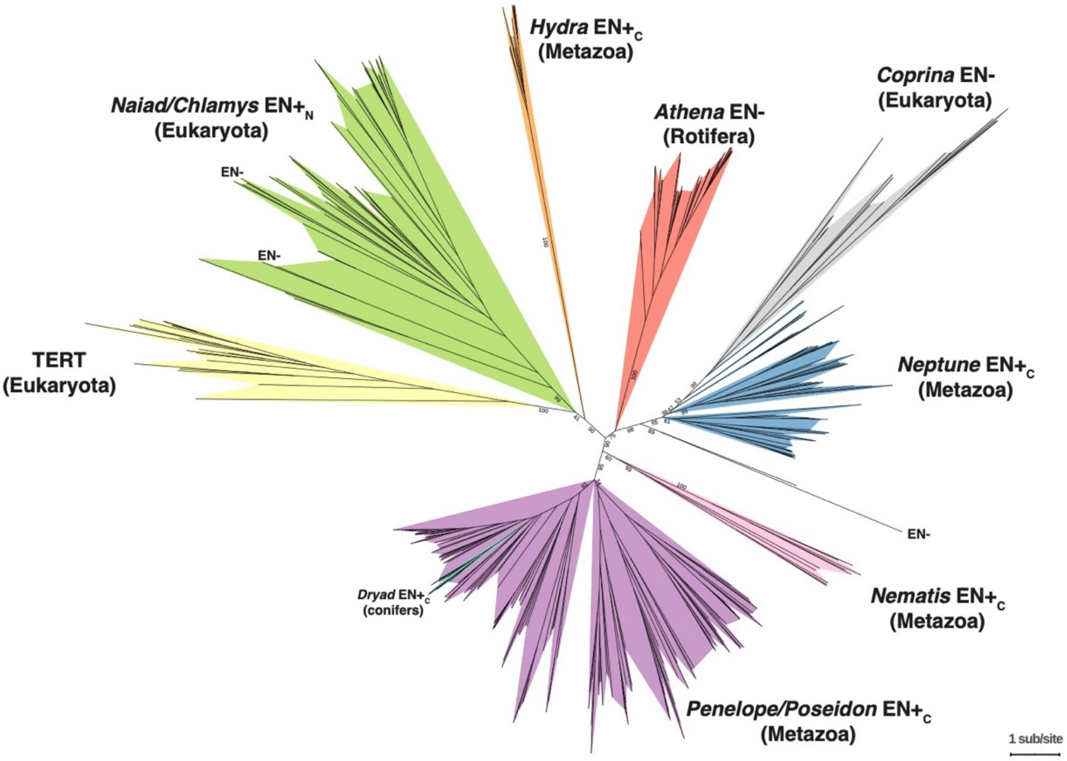


Figure 6. Reverse transcriptase phylogeny of PLEs and TERTs.
Phylogeny of canonical PLE superfamilies, *Hydra*, *Naiad/Chlamys* and TERTs based on core RT domain. Ultrafast bootstrap supports are displayed for key nodes.

Given the increased diversity of GIY-YIG ENs, I also attempted to further elucidate the evolutionary relationship of GIY-YIG ENs amongst PLEs. The domain is too short for typical phylogenetic analysis, and I therefore repeated the protein clustering approach of Dunin-Horkawicz et al. (2006) using GIY-YIG EN domains from the GIY-YIG superfamily annotated by NCBI (cd00719) and from all major groups of EN+ PLEs. The domains from PLEs formed a diffuse cluster, which was nonetheless clearly distinct from most other GIY-

YIG families (Figure 7). *Neptune* and *Penelope/Poseidon* formed distinct although strongly connected clusters, with the small number of *Nematis* sequences appearing to be intermediate between the two. This result is in line with the structure of the GIY-YIG ENs in these superfamilies, all of which share the CCHH motif, with *Neptune* containing the additional linker with $CX_{2-5}CxxC$ motif, *Nematis* containing a linker of reduced length with the $CX_{2-5}CxxC$ motif, and *Penelope/Poseidon* lacking the linker and $CX_{2-5}CxxC$ motif (Figure S1). *Naiad* was essentially indistinguishable from *Neptune,* in line with the presence of both the full-length linker/$CX_{2-5}CxxC$ motif and the CCHH motif. *Hydra* formed a well-resolved cluster distinct from other PLEs, in line with its unique linker motif and absence of the CCHH motif. *Chlamys* and especially the *Chlamys*-like elements were more diffusely clustered, most likely due to the presence of fewer well-conserved sites resulting from the lack of both the $CX_{2-5}CxxC$ and CCHH motifs. Recapturing the major result of Dunin-Horkawicz et al. (2006), the closest known sequences to the GIY-YIG EN of PLEs were from the HE_Tlr8p_PBC-V_like family. These GIY-YIG ENs are from homing endonucleases described from PBCV-1 and *Tlr8*, but also includes representatives from homing endonucleases of iridoviruses such as CIV and several bacteria. The HE_Tlr8p_PBC-V_like family formed a diffuse cluster, with several sequences appearing to be more strongly linked to domains from *Chlamys* and *Chlamys*-like elements than to other PLEs. In particular, domains from PBCV-1 (and other chloroviruses), iridoviruses and *Tlr8* (i.e. those that were originally linked to PLEs by Dunin-Horkawicz et al. (2006)) appeared to be linked to *PLE-2_cBra* and *PLE-5_cBra*, while bacterial domains were linked to *Chlamys*. These results should be interpreted very tentatively, since none of the HE_Tlr8p_PBC-V_like, *Chlamys* or *Chlamys*-like ENs formed well-resolved clusters.

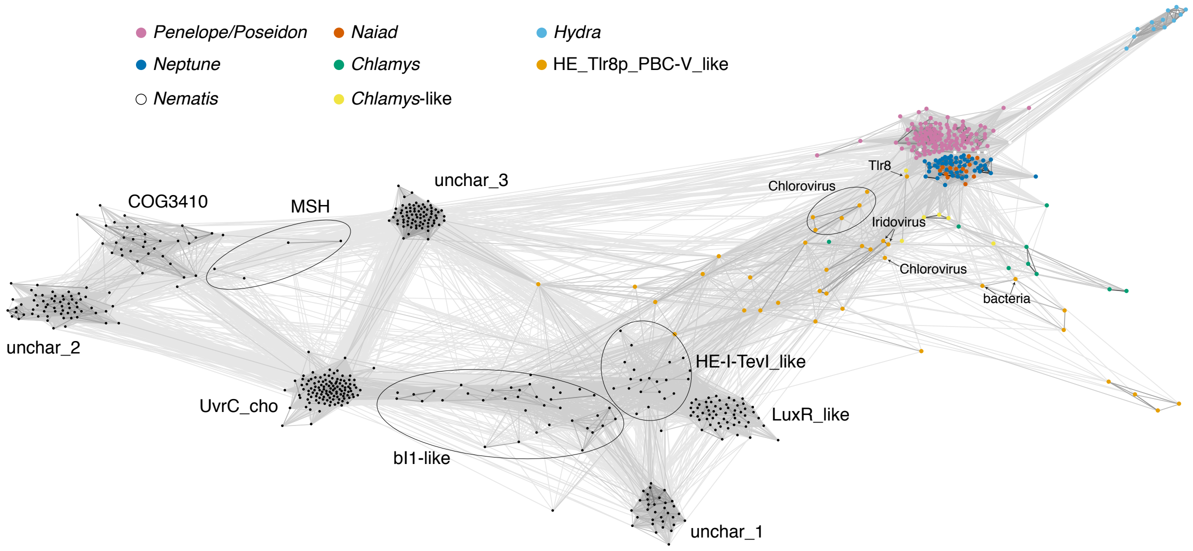


Figure 7. CLANS protein clustering of the GIY-YIG endonuclease domain.
GIY-YIG EN domains from PLEs and homing endonucleases of the HE_Tlr8p_PBC-V_like family are coloured, and particular HE_Tlr8p_PBC-V_like domains are highlighted. Clusters representing families from the NCBI GIY-YIG EN superfamily are highlighted.

5.5 Discussion

5.5.1 Diversity and recent activity of transposable elements in *Chlamydomonas reinhardtii*

Although an established TE library was available for *C. reinhardtii*, the completeness of these annotations was unclear. Via manual curation I have increased the proportion of the genome identified as TE sequence by ~50% and doubled the number of annotated TE families, demonstrating that although the original library included many of the most abundant TEs, it was only partially complete and substantially underestimated TE diversity. Furthermore, I have improved and extended the consensus sequences of existing models, amended several misclassifications, and split many existing families to the subfamily level. Given the exhaustive approach used, it is likely that the updated library is near complete. However, there are certain to be a small number of low-copy number TEs that have escaped annotation, especially if they are nonautonomous or do not encode recognised TE proteins (preventing identification based on protein domain searches). Furthermore, the completeness of the library is only directly relevant to the reference assembly in which the annotations were performed. Given the high activity of TEs and that several TEs were present at very low copy number (even single copy), it is likely that the genomes of other *C. reinhardtii* strains contain TEs that are not present in the reference genome. For example, *Pioneer1* was identified as an active element in the Florida isolate CC-2343 and was shown to be absent from laboratory strains (Graham et al. 1995). In a preliminary analysis of TEs in the CC-2931 assembly, I have curated autonomous *EnSpm* and *CryptonF* families that are actively transposing and are entirely absent from the reference genome (data not shown). Although they may comprise a small percentage of the total TE diversity in the species, it is possible that such elements are major components of the active repertoire of TEs in any particular strain, and care should be taken when performing species-wide analyses. Nonetheless, the library is expected to be a useful resource for *C. reinhardtii* research, with applications including the exploration of general features of genome architecture, incorporation into functional genomics analyses (e.g. methylation studies) and performing population genetics analyses of TE variants.

One of the notable results from the updated library is that the vast majority of TE copies exhibit minimal divergence from their consensus sequences, suggesting recent activity. This was supported qualitatively by comparison of the reference and CC-2931 assemblies, revealing insertion polymorphisms across all superfamilies. Transposable element landscapes such as that shown in Figure 1 are often used to reconstruct the evolutionary history of transposition in a species and its ancestors. Under the idealised assumption that inactive copies remain in the genome and that mutation occurs randomly both among TEs and in time, the divergence of individual TE copies from their respective consensus sequences can be used to infer what TEs were active at a given point, and in the case of copy and paste TEs, how much transposition was occurring. Such inferences must be interpreted carefully since these assumptions can very easily be violated to different degrees depending on the genome in question. Most obviously, older TEs would be expected to be underrepresented relative to their transposition rate at the time of their activity due to having had longer to be removed from the genome by deletion. Mutation is also unlikely to be random, although more sophisticated consensus callers that partially account for mutation bias have been developed (Storer et al. 2021). TE landscape plots should also only be created for species with high-quality TE annotations, since relying on consensus sequences produced in related species introduces bias in divergence estimates (Platt et al. 2016). Such approaches have most successfully been applied to species with larger genomes (e.g. vertebrates) where degraded TE copies have been retained over longer time periods and represent the majority of genomic TE sequence. TE landscapes for many animal genomes are available from the RepeatMasker website (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html).

The L-shaped distribution observed for *C. reinhardtii* could be explained under at least two scenarios. First, it is possible that the species has experienced a very low rate of transposition in the past and is now experiencing a burst of TE activity. This has been proposed as an explanation for the L-shaped TE landscape of *Drosophila melanogaster*, since the landscape in *Drosophila simulans* is comparatively flat, implying a lineage specific increase in activity in *D. melanogaster* (Mérel et al. 2020). Alternatively, the landscape may provide no information on the relative rates of current and past transposition if inactive copies are absent from the genome. This appears more likely to be the case in *C. reinhardtii*, with ~20% of TE sequence exhibiting >5% divergence from consensus (note that this proportion is considerably higher in *Drosophila*). Although TE landscapes will need to be produced for other *Chlamydomonas* species to specifically exclude a “burst” hypothesis, this is not supported by current evidence. The *C. reinhardtii* genome is ~20 Mb smaller than that of its closest relatives, which harbour more TE sequence (3.4.2). Furthermore, the families manually curated for *C. incerta* and *C. schloesseri* (3.4.2 and PLEs in this chapter) exhibited similarly low between copy divergence (data not shown), suggesting similar landscapes will be found across *Chlamydomonas*.

It is therefore possible that transposition rates have not differed greatly through time in *Chlamydomonas*, and that the L-shaped *C. reinhardtii* distribution can be explained by inactive TE sequence being efficiently removed from the genome (i.e. only recently active TEs are present and observed). One explanation for this would be a deletion bias. Ness et al. (2015) reported that deletions and insertions occurred at approximately similar rates in *C. reinhardtii* mutation accumulation (MA) lines, although deletions were significantly larger. This was based on short-read sequencing and focussed on short indels, and larger undetected events could contribute substantially to the overall balance of deletions and insertions. It remains to be seen if such a bias is strong enough to solely account for the efficient removal of inactive TE sequence, or if selection acting on genome size also plays a role. Although often considered a large genome for a unicellular species, the ~111 Mb *C. reinhardtii* genome is highly compact with respect to intergenic sequence and a large proportion of the genome size can be attributed to both the considerable number and length of introns (3.4.5, 3.4.9), which are generally not repetitive (4.4.3). It is possible that the *C. reinhardtii* genome is under strong selection for compactness, and that deletions of TE sequence are sufficiently advantageous to be efficiently fixed by selection.

The second remarkable feature of the *C. reinhardtii* annotation is the outstanding diversity of TEs. Following the discovery of PLEs and Cryptons, the species contains active elements from all major orders of TEs, except for the enigmatic Polintons. While this is of significance to studies of *C. reinhardtii*, it may be particularly pertinent in the wider context of green algal genomics. The chlorophyte lineage represents ~1 billion years of evolution (Leliaert et al. 2012), yet *C. reinhardtii* is the only species with anything approaching a complete TE library. With notable exceptions such as the annotation of a partial library in *V. carteri* (Prochnik et al. 2010) and the study of *Zepp* in *Coccomyxa subellipsoidea* (Blanc et al. 2012), there has been almost no curation of TEs in any chlorophyte species. The updated *C. reinhardtii* library is expected to be a useful resource for directly repeat masking closely related *Chlamydomonas* species and more generally for homology-based TE classification in more distantly related species, many of which are expected to have far more TE-rich genomes (e.g. Zhang et al. (2020)). Furthermore, the wealth of TE diversity raises several future questions about TE evolution in green algae. Do any of the DIRS elements described in *C. reinhardtii* represent a new superfamily, and are similar elements present across Chlorophyta? Have TET/JBP proteins been proliferated in algal genomes by Helitrons, and what effect has this had on the evolution of genome-wide methylation? How widespread are *CrpytonF* elements in chlorophytes, and have these been inherited vertically or horizontally from fungi or oomycetes? With the ever-expanding number of algal genome projects, it will soon be possible to address these questions and several more.

The inferred recent activity of such a diverse repertoire of TEs raises further questions about the evolution of *C. reinhardtii*. The species has presumably evolved considerable defence mechanisms to suppress TE activity and the potential mutagenic effects of transposition. TE silencing in *C. reinhardtii* is thought to occur at both the transcriptional and post-transcriptional level via a number of partly independent mechanisms (van Dijk et al. 2006). These include repressive histone modifications (Jeong et al. 2002; Zhang et al. 2002) and RNA interference (Casas-Mollano et al. 2008). With the possible exception of the most repetitive parts of the genome, CG methylation does not appear to play a general role in TE silencing (Lopez et al. 2015), although the discovery of TEs encoding TET-JBP proteins at least hints towards a relationship between methylation and TE regulation (4.4.8). A related question is to what extent have TEs evolved to minimise their mutagenic effects on the host genome? TEs such as *TOC1* and *Tcr1* were originally identified based on their insertion into genic sequence (Day et al. 1988; Schnell and Lefebvre 1993), although it is not clear if these were rare events. TEs in *C. reinhardtii* are clearly underrepresented in both exonic and intronic sequences (Philippsen et al. 2016), although such a pattern could solely be produced by selection acting against genic insertions. However, several families exhibited targeted insertions at microsatellites, which are generally non-exonic. These include a number of *Chlamys* elements and the LINE element *L1-5_cRei,* which targets repeats that are highly enriched in subtelomeres (Chaux-Jukic et al. 2021). As discussed in 5.4.2, a diverse set of families also encode PHD fingers, which may be involved in targeted insertion. It is of course expected that insertion biases, where they exist, differ by TE type and family, and that the observed genomic landscape of TE copies is likely shaped by both targeted insertion and selection acting against genic insertions. Fully addressing this question will require population genetics studies or the identification of transpositions in MA lines.

5.5.2 *Naiad/Chlamys* is a major new clade of retrotransposons

Previous studies described PLEs as an order of retrotransposons comprised of both EN– and C-terminal EN+ elements. This group, which I refer to as canonical PLEs, can be divided into the superfamilies *Penelope/Poseidon, Neptune* and *Nematis* (all EN+), and *Coprina* and *Athena* (both EN–). The work presented in this chapter shows that *Naiad/Chlamys* is a second major clade of PLEs that is clearly distinct from canonical PLEs. As with canonical PLEs, *Naiad/Chlamys* is found across several eukaryotic kingdoms and also contains EN– and EN+ elements, although the EN is located at the N-terminus. Most of the described diversity forms two EN+ superfamilies, the animal *Naiad* elements and the *Chlamys* elements of Viridiplantae. Remarkably, a functionally characterised *Chlamys* element appears to share the mechanistic features of *Penelope*, suggesting an ancient and deeply conserved functional organisation in PLEs.

Aside from the N-terminal GIY-YIG EN, one distinguishing feature of *Naiad/Chlamys* is the CxC domain upstream of the RT N-terminal region. This conserved motif does not have any matches in databases, although the conserved cysteines and histidine could suggest a zinc finger motif. The CX$_{2-5}$CxxC zinc finger present in the linker of the GIY-YIG EN of *Neptune* and *Nematis* elements has been hypothesised to play a role in microsatellite insertion biases (Arkhipova 2006). Such biases were observed for several *Chlamys* and *Naiad* families, with (C)$_n$ and (CA)$_n$ commonly observed in *Chlamys*, and (TA)$_n$ in *Naiad*. While *Naiad* also encode the CX$_{2-5}$CxxC motif in the GIY-YIG EN, *Chlamys* do not, and it is possible that the novel CxC domain plays a role in targeting EN activity to specific DNA repeats.

Many unanswered questions relate to the *Chlamys*-like elements found in *C. braunii*, two *Selaginella* species and *P. polycephalum*. Based on the current phylogenetic analyses, there are two independent groups of EN– elements, those found in *P. polycephalum* (which are also the only *Naiad/Chlamys* families lacking the CxC domain), and those found in the streptophyte species. Most interestingly, neither of these groups are basal to all EN+ elements, which if correct could imply that they have lost EN domains independently. Alternatively, there may have been multiple gains of EN, which as discussed below has almost certainly occurred in PLE evolution. Furthermore, unlike almost all described EN– canonical PLEs, none of the EN– *Chlamys*-like families were associated with telomeres. It is not yet clear how these families are able to retrotranspose genome-wide without EN. Of particular interest is *PLE-X1_sTam* of *S. tamarschina*, which to my knowledge is the first known TE to exhibit targeted insertion to the 28s ribosomal RNA gene in plants, as is observed in certain LINE elements in arthropods and other animals (Eickbush 2002; Penton and Crease 2004). Addressing these questions and attempting to expand the number of annotated *Chlamys*-like families will be a key aim of future PLE research. It is possible that at least some of these currently isolated families are representatives of substantial clades that are awaiting to be discovered. The sequencing of additional genomes from other charophytes and "early-diverging" plants, as well as other protists related to *P. polycephalum*, may be particularly useful in this endeavour.

Finally, it is interesting to consider the *Naiad* elements from the Chinese razor clam and the African social spider that appear to encode selenoproteins, which are noteworthy for two reasons. First, to my knowledge these are the first described instances of TEs encoding selenoproteins. Second, the vast majority of described selenoproteins encode a single selenocysteine, whereas both *Naiad-1_sCon* and *Naiad-1_sDum* encode four. Baclaocos et al. (2019) performed an analysis of selenoprotein P (SelP), one of the few selenoproteins encoding multiple selenocysteines, finding that in bivalves SelP contains the most selenocysteine residues of any metazoan group, and that in spiders SelP contains a moderate number of selenocysteines. Bivalves in particular are known for their high selenium content (Bryszewska and Måge 2015), and it may be that these two families represent cases of TEs adapting to their host cellular environments. However, even in bivalves selenoproteins are very rare (e.g. the selenoproteome of the pacific oyster contains only 32 genes (Baclaocos et al. 2019)), suggesting a more specific role for the replacement of cysteine by selenocysteine. Selenocysteine residues are involved in numerous physiological processes and are generally found at catalytic sites, where in many cases they may have a catalytic advantage relative to cysteine (Labunskyy et al. 2014). The selenocysteines in the *Naiad* peptides are mostly present at highly conserved sites in the CX$_{2-5}$CxxC, CCHH and DK motifs, and although the precise physiological roles of these motifs in PLEs is unknown, it may be that the incorporation of selenocysteine provides both a catalytic and evolutionary advantage. It remains to be seen if these represent highly unusual cases, although the fact that they appear to have evolved independently would suggest that several other *Naiad* elements encoding selenoproteins exist. This may have implications for TE annotation in general, and selenoprotein-encoding TEs may have previously been overlooked in taxa such as bivalves. Additionally, this result may provide insight into the evolution of new selenoproteins. The transition from encoding cysteine to selenocysteine is expected to be a complex evolutionary process, since a gene must acquire a SECIS element and near-simultaneously undergo a mutation from TGT/TGC (encoding Cys) to TGA (Castellano et al. 2004). The insertion of TEs carrying SECIS elements into the 3' UTRs of genes could provide a pathway for SECIS acquisition, especially for TEs that undergo 5' truncation and may insert with little additional sequence. It will be interesting to determine if the selenoprotein-encoding *Naiads*, or indeed any other TEs, have contributed to the evolution of new selenoproteins in their host genomes.

5.5.3 The evolution of *Penelope*-like elements

The discovery of *Naiad/Chlamys* substantially develops our understanding of PLE evolution. It has been hypothesised that the common ancestor of PLEs and TERTs may have resembled *Athena* and *Coprina* elements (Gladyshev and Arkhipova 2007). Such an EN–, telomere-associated retroelement could have been co-opted in early eukaryotic evolution to actively maintain telomeric sequence, leading to the evolution of TERTs. Similar elements could have acquired C-terminal GIY-YIG EN, leading to the evolution of canonical PLEs capable of genome-wide insertion. Based on parsimony, it has been suggested that the gain of the GIY-YIG EN in canonical PLEs may have occurred once (Arkhipova 2006). In that case, the most likely scenario would be that the *Neptune* GIY-YIG with CCHH and linker/CX$_{2-5}$CxxC motifs is ancestral, with the linker/CX$_{2-5}$CxxC being reduced in *Nematis*, and entirely lost in *Penelope/Poseidon*. Recent phylogenetic analyses have, however, suggested a more complex view. While *Athena* do appear to branch basally to EN+ PLEs, *Coprina* appears to form a clade with *Neptune* (Arkhipova et al. 2017). Following this, *Neptune* does not form a well-supported clade with *Nematis* + *Penelope/Poseidon*. The complexity of the evolution of PLEs is underlined by the discovery of EN– *Neptune*-like elements such as *MjPLE01* from the kuruma shrimp *Marsupenaeus japonicus* (Koyama et al. 2013). It therefore appears that even within the evolution of canonical PLEs, EN may have been gained and lost more than once.

The IFD insertion observed in most *Naiad/Chlamys* elements potentially strengthens the link between PLEs and TERTs. Thought to be unique to TERTs, it is possible that a shorter precursor of the IFDb/TRAP domain, as found in *Naiad/Chlamys*, was present in the common ancestor of PLEs and TERTs, where it later became extended in TERTs and lost in canonical PLEs. In functional TERTs, the TRAP domain forms a structure that aids the stabilisation of telomerase RNA and DNA during the extension of telomeric DNA (Jiang et al. 2018). It is unclear what the role of the insertion may be in *Naiad/Chlamys*, and it could alternatively be an independent insertion that has occurred specifically in this group of PLEs. The fact that it has been lost in certain *Naiad* elements demonstrates that it is not necessarily a functional requirement. The evolution of the GIY-YIG EN and its associated domains in *Naiad/Chlamys* and canonical EN+ PLEs is also complex. Although tentative, the *Chlamys* and *Chlamys*-like GIY-YIG ENs, which lack the CX$_{2-5}$CxxC and CCHH motifs, are potentially more similar to those from the HE_Tlr8p_PBC-V_like family, which are mostly found in homing endonucleases. It is possible that the CX$_{2-5}$CxxC and CCHH motifs are derived, and that the GIY-YIG ENs of *Chlamys* and *Chlamys*-like elements more closely resemble the ancestral PLE GIY-YIG EN domain, which was potentially acquired from a homing endonuclease. This is supported by the phylogeny of *Naiad/Chlamys*, with the CX$_{2-5}$CxxC and CCHH motifs found only in *Naiad.* Finally, except for two families, HHR motifs were absent from *Naiad*/*Chlamys*, but present in most canonical PLEs and *Hydra*. The two *Naiad/Chlamys* elements with HHR are phylogenetically distant from each other and they may have acquired these motifs independently. HHRs have been hypothesised to self-cleave pre-cursor RNA to give rise to compatible 5' and 3' ends that could form a circular RNA template for retrotranscription (Cervera and De la Peña 2014). However, the ribozyme cleavage site does not coincide with the TE-host boundary, and the role of HRRs in PLE retrotransposition remains unclear (Arkhipova et al. 2017).

Considering the diversity of PLEs collectively, several plausible scenarios of PLE evolution can be envisioned. It is possible that the ancestor of PLEs and TERTs was a telomere-associated retroelement similar to contemporary *Athena* and *Coprina* elements. This element may have contained an IFD insertion similar to those observed in *Naiad/Chlamys*, that was extended to form the TRAP domain in TERT evolution. Such an early element may have acquired an N-terminal GIY-YIG EN, most likely lacking $CX_{2-5}CxxC$ and CCHH motifs and potentially derived from a homing endonuclease, giving rise to the *Naiad/Chlamys* clade of PLEs. The EN may have been secondarily lost from certain enigmatic lineages of *Chlamys*-like elements that appear to have an alternative means of genome-wide insertion. The $CX_{2-5}CxxC$ and CCHH motifs may have evolved in *Naiad*, and been transferred multiple time to the C-terminus of distantly related EN– elements (that had potentially lost IFD insertions), giving rise to *Neptune*, *Nematis* + *Penelope/Poseidon* (which later underwent reduction and complete loss of the linker/$CX_{2-5}CxxC$ motif, respectively), and *Hydra* (which lost the CCHH motif and evolved an altered linker). Alternatively, the $CX_{2-5}CxxC$ and CCHH motifs could have evolved in canonical PLEs and been transferred to *Naiad*, replacing the ancestral domain found in other *Naiad/Chlamys* elements. As all of the PLEs containing $CX_{2-5}CxxC$ and CCHH motifs are found in animals, it is plausible that such EN shuffling occurred early in animal evolution. Indeed, the genomes of early animals must have contained a high diversity of PLEs, a situation that is still observed in certain species such as the coral *A. millepora*, which contains *Neptune, Naiad* and *Hydra* elements. HHR motifs may have evolved once in the ancestor of canonical PLEs and *Hydra*, although there is currently very limited phylogenetic support for such a clade. Various other permutations of these events could have produced the described diversity of elements, but what is clear is that the GIY-YIG EN has likely been acquired multiple times, producing at least four distinct groups exhibiting genome-wide retrotransposition (*Naiad/Chlamys, Neptune, Nematis + Penelope/Poseidon*, and *Hydra*).

Although it may never be possible to entirely recover the evolutionary history of PLEs, these results have important consequences for TE classification. PLEs are generally understudied and are largely overlooked in several classification systems. In the Dfam database (https://dfam.org/home) there are no superfamily-level classifications for PLEs, while in Repbase all PLEs are treated as a clade of non-LTR (i.e. LINE) elements (although it is recognised that this is not phylogenetically meaningful (Kojima 2020)). Such a neglected position likely stems from two primary reasons. First, PLEs are absent from well-studied species such as mammals, birds and angiosperms, and are potentially deemed to be more phylogenetically restricted than other TE orders. Second, PLEs are not considered to be as diverse as LINEs or LTRs, which each comprise ancient clades varying in their structural organisation of domains. I have demonstrated that these two assumptions are unfounded. EN+ PLEs are now documented from invertebrate and vertebrate animals, chlorophyte and streptophyte green algae, a small number of land plants (Lin et al. 2016) and the slime mold *P. physarum*. EN– PLEs are found in animals, plants and streptophyte green algae, fungi, red algae, heterokonts and *P. physarum*. Three primary functional organisations exist: EN–, C-terminal EN+, and N-terminal EN+. These groups can each be further split into deep branching clades or superfamilies, each of which possess distinguishing structural features and domains. Additionally, as with SINEs parasitising LINEs, the newly described retrozymes likely parasitise autonomous PLEs in animals (Cervera and de la Peña 2020). As more genomes from neglected and phylogenetically diverse lineages become available it is likely that the diversity of PLEs will continue to expand, further supporting their increasingly important and unique position in TE biology.

5.6 Methods

5.6.1 Curation of transposable elements in *Chlamydomonas*

TE curation was performed following the methodology documented in 1.4.3. Briefly, multiple copies of a putative TEs were collected, aligned, visualised and summarised as consensus sequences. TEs were classified based on nucleotide and protein (if autonomous) homology and characteristic structures and motifs (LTRs, TIRs, 5' truncation, TSDs, etc.). Where relevant, TEs were curated to the subfamily level. Annotation was initially performed using the v5 assembly, although all consensus sequences were later checked against the CC-503 v6 assembly and updated if necessary. In many cases extrinsic evidence supporting precise TE boundaries was provided by observing polymorphic insertions between the reference assembly and CC-2931 assembly (5.6.4). Iso-Seq data were used to identify gene models and putative peptides in TEs containing transcribed genes with introns.

Preliminary repeat models used for curation were produced by running RepeatModeler v1.0.11 (Smit and Hubley 2008-2015) on the *C. reinhardtii* v5 reference assembly. Each model was then curated as described above, with models that did not appear to be TEs discarded. All existing *C. reinhardtii* TE sequences from Repbase and the literature were also curated using the same approach. Several low-copy number TEs were identified by attempting to curate genic regions encoding TE-related protein domains in the v5.6 gene annotations. A small number of single-copy TEs were identified based only on a TE-related protein domain and a clean insertion polymorphism relative to CC-2931 (as well as any structural motifs, if present).

All curated TEs were given unique names following standard nomenclature guidelines. The structure "*superfamily-X_cRei*" was used, where superfamily is the element superfamily (e.g. *L1*, *Gypsy*, *hAT*) and X represents a unique number. Nonautonomous elements were designated by the inclusion of an additional "N" (e.g. *EnSpm-N1_cRei*, *RTEX-N3_cRei*) and elements with subfamilies were split alphanumerically (e.g. *Dualen-4a_cRei*, *Dualen-4b_cRei*). For elements not classified to the superfamily level the order was used in its place (e.g. LINE, LTR, DNA), and for unclassified elements "unknown" was used. In cases where TEs had existing names, the novel consensus sequences were nonetheless given new identifiers, which were listed as synonyms (Dataset S1). This was done to differentiate the updated and original consensus sequences, and additionally to resolve historic cases of redundancy (e.g. there are two entirely separate TEs that have been named *TOC2* (Day 1995; Goodwin and Poulter 2004)).

The genomic coordinates of TE sequences and the divergence of each TE sequence from its consensus sequence were identified by running RepeatMasker v4.0.9 (Smit et al. 2013-2015) on the CC-1690 assembly, using the updated consensus sequences as a custom library. CC-1690 was chosen since it is the most-contiguous assembly and both the CC-503 v6 and CC-4532 have been affected by structural mutations (large deletions in CC-503 and TE proliferation in CC-4532, 4.4.4, 4.4.6).

5.6.2 Curation of *Naiad/Chlamys* elements in other species

All autonomous and nonautonomous *C. reinhardtii* PLEs were curated as part of the library update. PLEs from *C. incerta, C. schloesseri* and *E. debaryana* were curated as part of their own library annotations (3.6.5). Using the identified protein sequences from *C. reinhardtii*, related PLEs were identified using a combination of PSI-BLAST and tblastn. PSI-BLAST was performed using NCBI servers, while tblastn was performed against all eukaryotic genome assemblies accessed from NCBI on 09/04/20. Genomes with multiple hits were selected for further curation, and where several closely related species had multiple hits the most contiguous assembly was chosen. A custom script was used to collect the nucleotide sequence of all tblastn hits in a given genome, which were then each queried against the genome using blastn to estimate copy number. The most abundant putative PLEs were then targeted for manual curation from alignments of multiple copies as described previously (1.4.3, 5.6.1).

Novel *Hydra* elements were identified following the same approach, using the existing proteins from Repbase as query sequences.

5.6.3 Phylogenetic and protein clustering analyses

Peptide sequences were identified from PLE consensus sequences based on translation of the longest ORF. Protein alignments were produced using MAFFT v7.273 (Katoh and Standley 2013) with the parameters "--genafpair" and --maxiterate 10000". For the phylogenetic analysis of *Naiad/Chlamys* (Figure 2), protein regions from the CxC domain to the RT thumb (see Figure 4) were aligned, and additional *Athena* core RT/thumb sequences were included as an outgroup. For the phylogenetic analysis of PLEs and TERTs (Figure 6) only the DKG domain, core RT and RT thumb were aligned. PLE and TERT protein sequences were obtained from Repbase, Gladyshev and Arkhipova (2007) and Lin et al. (2016). Phylogenies were produced using IQ-TREE v1.6.9 (Nguyen et al. 2015), run with ModelFinder ("-m MFP") (Kalyaanamoorthy et al. 2017). For the *Naiad/Chlamys* phylogeny, conventional bootstrapping was performed ("-B 1000"), while for the PLE/TERT phylogeny, ultrafast bootstrapping was performed ("-bb 1000") (Hoang et al. 2018). One of the major differences in the interpretation of these support values is that ultrafast bootstrapping is relatively unbiased, so that >95% should be considered as strong support (as opposed to >70%, which is typically considered as strong support in conventional bootstrapping).

Protein clustering analysis of the GIY-YIG EN was performed using CLANS (Frickey and Lupas 2004). GIY-YIG ENs from all GIY-YIG superfamilies annotated at NCBI (cd00719) were combined with those from PLEs. GIY-YIG ENs from NCBI were truncated from the

“GIY” motif to the conserved asparagine. PLE GIY-YIG ENs were truncated from the CX$_{2-5}$CxxC motif (if present) to the conserved asparagine (see Figure S1). After a preliminary analysis, several families that were very weakly linked to PLEs were discarded. CLANS was run with a p-value threshold of 1×10^{-8} until no further changes were observed to clustering.

5.6.4 Genome assembly of CC-2931

High molecular weight DNA was extracted from a four-day culture of CC-2931 following the protocol documented introduced in 3.6.1 (Appendix C, Note S1). Genomic DNA was sequenced on the PacBio Sequel platform at Edinburgh Genomics, yielding 7.51 Gb of reads with an N50 of 20.96 kb. A *de novo* assembly was produced using wtdbg2 (Ruan and Li 2020) using the parameters “-g 111m” and “-x sq”. PacBio-based polishing was performed with two iterations of Arrow (https://github.com/PacificBiosciences/GenomicConsensus), mapping reads with pbmm2 (https://github.com/PacificBiosciences/pbmm2). Illumina-based polishing was performed with one iteration of Pilon (Walker et al. 2014) with the flag “--fix bases”. Illumina data were obtained by the whole-genome re-sequencing of 14 MA lines derived from CC-2931 (Ness et al. 2015). Data from each line were subsampled to 10% coverage to ensure that no mutations, which are expected to be unique to single MA lines, were incorporated into the assembly. The assembly spanned 108.95 Mb on 177 contigs with an N50 of 3.01 Mb.

The polished *de novo* assembled contigs were manually scaffolded to chromosomes by alignment to the CC-1690 assembly using MashMap v2.0 (Jain et al. 2018a). Gaps between contigs were filled with 10 kb of “N” unknown bases. All regions where a contig of the CC-2931 assembly consistently mapped across two chromosomes of the CC-1690 assembly were visually inspected relative to the PacBio reads using IGV v2.7.2 (Robinson et al. 2011). All such breaks in synteny were supported by the raw reads, implying that there are three reciprocal translocations in the CC-2931 genome relative to laboratory strains (Figure S6). These putative rearrangements require further investigation before they can be confirmed, and the current chromosomal assembly is preliminary. The chromosomal assembly consisted of 17 chromosomes and 125 unplaced contigs, with ~98% of sequence assembled on chromosomes. TE polymorphisms were observed in IGV after mapping the chromosomal CC-2931 assembly to the relevant reference assembly using minimap2 v2.17 (Li 2018) with the parameter “-ax asm10”.

Chapter 6

General Discussion

6.1 Thesis Overview

The overarching aim of this thesis was to enhance our understanding of the evolutionary genomics of *Chlamydomonas reinhardtii* and its close relatives. In the process, I aimed to produce several resources that will facilitate the continued development of *C. reinhardtii* and *Chlamydomonas* as study systems for evolutionary research. *C. reinhardtii* has generally been studied in isolation, and one of the primary objectives was to provide a general framework for evolutionary research, from both population and between-species perspectives.

In Chapter 2, I used whole-genome re-sequencing data from all confirmed field isolates of *C. reinhardtii* to explore the demography and ecology of the species. I found that *C. reinhardtii* forms three highly differentiated and geographically structured lineages based on current sampling. I reported evidence of admixture and potentially migration between the two lineages present in North America. I found that field isolates from Quebec, the only site with multiple sampled individuals, showed little evidence of either spatial or temporal population structure. However, I found that the genomes of pairs of individuals shared large identical by descent haplotypes at far higher frequencies than would be expected. This result remains unexplained and may have implications for other microbial eukaryotes with similar life cycles. The work in this chapter confirms that the Quebec isolates currently represent the best available sample for population genetics analyses. It also lays the groundwork for developing *C. reinhardtii* as a model to study the evolutionary ecology of microbial eukaryotes.

In Chapter 3, I sequenced, assembled and annotated high quality genome assemblies for the two closest known relatives of *C. reinhardtii*, *Chlamydomonas incerta* and *Chlamydomonas schloesseri*, and one more distantly related unicellular species, *Edaphochlamys debaryana.* I characterised patterns of synteny between the *Chlamydomonas* genomes, finding limited evidence for large scale rearrangements. I described the major centromeric repeat in *C. reinhardtii* and found that centromeres and several other features of genome architecture are likely conserved between the *Chlamydomonas* species. I used patterns of nucleotide divergence across *Chlamydomonas* and more distantly related species to identify putative false positive and novel genes in *C. reinhardtii.* I also identified evolutionarily conserved elements and reported that longer introns do not contain a higher proportion of conserved sites. The work in this chapter presents resources that enable comparative genomics and molecular evolution analyses to be performed.

In Chapter 4, I detail my role in updating the *C. reinhardtii* reference genome assembly and annotation. The contiguity of the assembly was increased by an order of magnitude and several large misassemblies were fixed. The assembly improvements resulted in a corresponding improvement in gene annotation, since a substantial number of the gaps in previous assembly versions were in genic regions. More than 1,000 transposable element (TE) genes were also removed from the main annotation. Comparing the genome assemblies of three laboratory strains revealed the presence of several structural mutations and transposition events, many of which have disrupted genes. The *C. reinhardtii* reference assembly and annotation are fundamental to almost all aspects of evolutionary genomics research in *Chlamydomonas*.

Finally, in Chapter 5, I manually curated TE sequences in *C. reinhardtii* and produced a near complete TE library for the species. I reported that the *C. reinhardtii* genome harbours an unusually high diversity of TEs and that almost all TEs show evidence of recent activity. I described a major new clade of *Penelope*-like elements (PLEs) based on the TEs identified in *Chlamydomonas*. The work in this chapter enables us to study the population genetics and comparative genomics of TEs in *Chlamydomonas* in more detail than has previously been possible. The annotations will also enable us to characterise transposition events in mutation accumulation lines. Furthermore, the discovery of an entirely undescribed clade of PLEs demonstrates the potential benefits of studying species that are phylogenetically distant from typical model organisms.

In the following sections, I summarise the current state of evolutionary genomics resources in *Chlamydomonas*, focussing on the reference genome, gene annotations, and the datasets available for population genetics and comparative genomics analyses. In each instance, I provide context by briefly drawing comparisons to other model organisms. Finally, I discuss how we could tackle the issues in sampling new isolates of *Chlamydomonas*, which presents one of the largest obstacles to the continued use of the species in evolutionary research.

6.2 Evolutionary Genomics Resources for *Chlamydomonas*

6.2.1 The *Chlamydomonas reinhardtii* reference genome

A consistent and high-quality reference assembly is the starting point for almost all evolutionary genomics analyses. For the foreseeable future, nearly all research in *C. reinhardtii* is expected to utilise the CC-4532 v6 assembly (4.4.5), likely supplemented with the mating type *plus* locus of CC-503 v6, and the organelle genome assemblies of Gallaher et al. (2018). The more contiguous, but unannotated, CC-1690 assembly (O'Donnell et al. 2020) will also prove useful for any specific analyses focussing on the most repetitive parts of the genome, such as centromeres (4.4.3) and subtelomeres (Chaux-Jukic et al. 2021).

One of the most important improvements between the v5 assembly and CC-4532 v6 is that all genomic sequence is now expected to be ordered and orientated on chromosomes correctly.

This is important for any genome-wide analyses involving recombination or experimental designs involving crosses, several of which are currently planned or in process. The assembly of the biologically correct karyotype will also be important for comparative genomics analyses, including the characterisation of chromosomal rearrangements within and between-species. CC-4532 v6 is expected to be almost complete with respect to genic sequence, providing a comprehensive view of all but the most repetitive genomic regions. As detailed in 4.4.7, the CC-4532 genome does contain a substantial number of derived TE insertions. Although this is not ideal, the majority of these TE insertions are in intergenic regions and they can easily be accounted for or ignored in most analyses. Overall, the genome is expected to be an excellent resource for read mapping, variant calling and general comparative genomics analyses.

These improvements have raised the *C. reinhardtii* reference genome to a level approaching that of other model organisms with similarly complex genomes. However, there are two important points to note in this respect. First, the genomes of species such as *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) and *Caenorhabditis elegans* (C. elegans Sequencing Consortium 1998; Hillier et al. 2005) were essentially completed using Sanger sequencing at a time when the *C. reinhardtii* genome had not been assembled onto chromosomes. The genome assembly of *Drosophila melanogaster* has undergone a more continuous development with several major updates (Hoskins et al. 2007; Hoskins et al. 2015), although this genome is more repetitive and contains challenging heterochromatic regions. Second, with the development of long read sequencing, reference-quality genome assemblies for these model organisms have now progressed beyond single individuals, as demonstrated by the recent chromosome-level assemblies of non-reference individuals of *A. thaliana* (Jiao and Schneeberger 2020) and *D. melanogaster* (Adams et al. 2020). These resources enable the discovery of major genomic rearrangements and copy number variants that segregate within-species, and allow researchers to characterise any biases that may be introduced by performing analyses relative to a single reference. Although not discussed in any detail in this thesis, the chromosome-level assembly produced for the North Carolina field isolate CC-2931 in Chapter 5 is a first step towards developing similar resources for *C. reinhardtii*. The assemblies produced in this thesis suggest that it is now possible (and relatively affordable) to assemble chromosome-level assemblies for *C. reinhardtii* using existing technologies, and we can expect to see genomes for several additional strains assembled in the near future.

6.2.2 The *Chlamydomonas reinhardtii* structural annotations

The CC-4532 v6.1 annotation is a major improvement on v5.6 (4.4.7) and is expected to be sufficient for most evolutionary analyses, especially for the many questions that consider protein-coding genes collectively. For example, in the future we may want to study variation in certain evolutionary phenomena (recombination rate, mutation spectra, etc.) with respect to different genomic site classes (coding sequence, UTRs, introns, etc.), and the expected quality of the annotation should provide the platform to perform such analyses to a very high standard. One key development was the removal of over 1,000 TE genes from v5.6 (4.4.8), the presence of which would be expected to confound several analyses, including the example just given. To provide a second example, it would be interesting to characterise the genomic distribution of TE insertions and the frequency of these insertion variants in a population sample. We could then ask if certain categories of TE insertions, such as those disrupting exons, appear to be more deleterious than insertions in other genomic regions. However, to do this successfully it would be critical to have a clear distinction between TEs and standard protein-coding genes, as we have now achieved. The removal of several hundred protein-coding genes that were likely annotation artefacts is also expected to improve the resolution of genome-wide analyses.

Although the underlying gene models are expected to be accurate, the current limitations are more focussed on functional annotation, and there remain a substantial number of lineage-specific genes in *C. reinhardtii* of unknown function (Blaby and Blaby-Haas 2017). There also remains considerable scope to validate and potentially improve specific aspects of the annotation, for example the large number of newly predicted alternative isoforms (4.4.7), which have not been widely studied in the species (although see Labadorf et al. (2010) and Raj-Kumar et al. (2017)). Furthermore, the annotations are entirely limited to protein-coding genes, so for example if one wanted to characterise selection acting on long noncoding RNA genes (e.g. Wiberg et al. (2015)), this would not currently be possible. These shortfalls are expected to limit the evolutionary inferences that can be drawn from particular analyses, for example it is often informative to incorporate functional annotation (e.g. gene ontology terms) to further characterise gene sets delineated by certain analyses (e.g. a screen for positive selection). Furthermore, although the *C. reinhardtii* TE library I have produced is expected to be comprehensive, the inclusion of annotated TE genes in CC-4532 v6.1 (4.4.8) is certainly not complete and would require careful manual annotation to develop further.

The differences in gene annotation quality between *C. reinhardtii* and other model species are far starker than those between the current genome assemblies (6.2.1). For example, the recent Araport11 version of the *A. thaliana* genome introduced several thousand noncoding RNA genes and included detailed analysis of alternative splicing (Cheng et al. 2017). *D. melanogaster* is one of the best annotated genomes of any species, with nearly all coding and noncoding genes having undergone manual curation (Matthews et al. 2015), which has resulted in the discovery of a wide diversity of non-canonical genes (Crosby et al. 2015). These model systems also typically have their own dedicated genome hubs (e.g. TAIR, FlyBase), which provide a wealth of functional information beyond that which is currently available for *C. reinhardtii* from Phytozome. However, it should be noted that the significance of these comparisons is limited, since highly developed annotations are generally not as relevant for evolutionary analyses as they are for many other fields. As outlined above, for the majority of analyses the current gene models are expected to provide an accurate and informative resource. Nonetheless, further developing the *C. reinhardtii* gene annotations will be an important goal for the wider *Chlamydomonas* research community.

6.2.3 Population genetics resources

As outlined in 1.2, *C. reinhardtii* has almost exclusively been studied from the perspective of a single line of related laboratory strains. We have now produced or collated whole-genome re-sequencing data from all 36 known field isolates of the species. Of these, 25 were sampled from two fields ~80 km apart in Quebec. These isolates exhibit little signature of population structure and currently represent the best available sample for performing general population genetics analyses in *C. reinhardtii*. A subset of these isolates was used to calculate genetic diversity for particular analyses in Chapters 3 and 4, and has been used elsewhere to characterise patterns of linkage disequilibrium (LD) both genome-wide (Hasan and Ness 2020) and specifically at the mating type locus (Hasan et al. 2019). These isolates are expected to form the basis of many future population genetics analyses, for example they would provide a suitable population in which to identify TE variants and estimate the site frequency spectrum of TE insertions, as suggested in 2.6.2. However, we still understand very little about patterns of genetic diversity in the *C. reinhardtii* genome, and attempting to explain the unusual extent of haplotype sharing among the Quebec isolates will be a priority. Indeed, extensive haplotype sharing does not appear to be unique to the Quebec sample, and understanding the evolutionary forces and population processes that shape this result will be a general aim of future population genetics research in *C. reinhardtii*.

Although the Quebec isolates represent one of the very few population samples available for microbial eukaryotes, there are several interesting questions that could only be addressed by vastly increasing sampling. Fundamentally, we have a very rudimentary and likely highly biased view of the global distribution and evolutionary history of *C. reinhardtii*. The results presented in Chapter 2 hint at the existence of reproductive isolation between the two differentiated lineages present in N. America, and if more samples could be collected, it may be possible to extend our knowledge of how *C. reinhardtii* is genetically structured to far wider geographic scales, and even potentially to develop the species as a model to study the evolutionary ecology of microbial eukaryotes. This has been achieved in yeast species, where over 1,000 isolates of *Saccharomyces cerevisiae* were recently sequenced in a large-scale analysis investigating patterns of global population structure (Peter et al. 2018). The domestication of *S. cerevisiae* complicates its evolutionary history, although more than 300 isolates have now been sequenced from the wild yeast species complex *Saccharomyces paradoxus*, providing unprecedented insights into migration, reproductive isolation and speciation in microbial eukaryotes (Leducq et al. 2016; Eberlein et al. 2019).

Beyond microbial species, very large population datasets have now been collected and sequenced for several model organisms, including more than 1,300 *A. thaliana* individuals (The 1001 Genomes Consortium 2016) and more than 600 *D. melanogaster* individuals (Lack et al. 2015). As well as forming outstanding resources for addressing fundamental population genetics questions, such extensive sampling provides opportunities to study a wide array of additional population processes, including demographic history, admixture and introgression, and local adaptation. These samples are also excellent resources for performing many quantitative genetics analyses (e.g. genome-wide association studies). Although we largely lack the ecological or phenotypic knowledge of *C. reinhardtii* field isolates that would inform such analyses, there are nonetheless several interesting questions that could be asked if far larger samples could be obtained. For example, given that we expect sex to be induced by poor conditions, we might expect that the rate of sex in local populations is strongly influenced by environmental factors. With sampling of many populations from different environments, it would be possible to explore patterns of genetic diversity and LD between populations to address this question. Furthermore, given that the known range of the species extends from Florida to Quebec, we may also expect local adaptation to be prevalent. Although it is beyond the scope of this work, studying a far wider sample of field isolates may also provide opportunities for uncovering gene function (either experimentally or via genome association studies), or even for the selective breeding of traits of interest (e.g. biofuel yield), an approach that has been neglected in algal biotechnology.

Finally, as with the example of *S. cerevisiae* and *S. paradoxus*, several population genetics datasets have been produced for close relatives of model organisms. These include the closest relatives of *D. melanogaster*, *Drosophila simulans* and *Drosophila yakuba* (Rogers et al. 2014; Jackson et al. 2017), and relatives of *A. thaliana* in the Brassicaceae such as *Capsella grandiflora* (Williamson et al. 2014; Steige et al. 2017). The ability to study population genetics processes in multiple close relatives can provide insights into how different evolutionary histories or changes in life history traits can influence patterns of genetic diversity and genome evolution. Unfortunately, the sampling of additional *C. incerta* and *C. schloesseri* isolates is severely limited (3.4.1), and I am unaware of the existence of any other whole-genome population genetics datasets in the entirety of green algae (>1 billion years of evolution).

6.2.4 Comparative genomics resources

Prior to the work presented in this thesis, there were no genomics resources available for any close relatives of *C. reinhardtii*. Comparative genomics resources are required to perform several analyses, including molecular evolution analyses, the use of signatures of nucleotide divergence to identify coding and functional noncoding sequences, and the general study of genome evolution (genome size, repeat content, karyotype, etc.) between species. More broadly, comparative genomics resources can be thought of as necessary to place the overall biology of *C. reinhardtii* in an evolutionary context. The genomes assembled for *C. incerta, C. schloesseri,* and to a lesser extent *E. debaryana*, represent the first step towards performing such analyses for *C. reinhardtii*. However, the comparative genomics resources available for *Chlamydomonas* still compare poorly to several other model taxa and many analyses could be substantially improved by the sequencing of additional species.

The availability of one or more closely related outgroups is required for several analyses in population genetics and molecular evolution. For example, estimating divergence at putatively neutral and selected sites underlies the McDonald-Kreitman test for positive selection (McDonald and Kreitman 1991), and several extensions of this test have been developed that are based on polarising polymorphisms by comparison to an outgroup (or outgroups) and obtaining the unfolded site frequency spectrum (Booker et al. 2017). Since ancestral polymorphism contributes to divergence estimates, optimal outgroups are sufficiently divergent from the focal species, such that lineage sorting has mostly been completed. Mugal et al. (2020) provided guidelines for choosing an appropriate outgroup, recommending that t (time in coalescent units) should be greater than 5 following the equation $t = D/(2\theta) - 1$ (where D is divergence and θ is genetic diversity in the focal species). Taking the estimate of genetic diversity for the Quebec population at four-fold degenerate (4D) sites of 0.0236 (2.4.6) and the estimate of divergence between *C. reinhardtii* and *C. incerta* at 4D sites of 0.34 (3.4.7), $t \approx 6.2$, suggesting that *C. incerta* is a suitable outgroup for future analyses. However, it is unlikely that 4D sites are evolving neutrally in *Chlamydomonas* (Popescu et al. 2006), and a more appropriate neutral reference is required. As a result of selection acting on codon usage, short introns are frequently used as a neutral reference class in *Drosophila* population genetics analyses (Halligan and Keightley 2006; Parsch et al. 2010). Given their abundance and general lack of constraint in the *C. reinhardtii* genome (3.4.9), introns may also be an attractive putatively neutral reference in *Chlamydomonas*. However, less than 50% of intronic sites can be aligned between *C. reinhardtii* and *C. incerta* (3.4.7) and the high frequency of indels in these regions complicates their use as reference sequences to estimate putatively neutral divergence (Rob Ness, unpublished). Thus, a more closely related outgroup would be desirable, if such a species exists.

Genome assemblies for appropriate outgroup species have now been produced for many model organisms. For example, the genomes of *D. simulans* and *D. yakuba* (Drosophila 12 Genomes Consortium 2007) have frequently been used as outgroups in analyses of selection in *D. melanogaster* (e.g. Keightley et al. (2016)). For the aforementioned example of *C. grandiflora*, divergence estimates to *A. thaliana* have been used (Williamson et al. 2014; Steige et al. 2017). In many respects, the current situation for *C. reinhardtii* and *Chlamydomonas* mirrors the past situation of *C. elegans* and *Caenorhabditis*. For most of its history as a model organism, *C. elegans* was studied in phylogenetic isolation due to a lack of sampled relatives (and a corresponding lack of the ecological knowledge to perform sampling). A genome assembly for *Caenorhabditis briggsae* was assembled to enable comparative genomics analyses, although average synonymous divergence between *C. elegans* and *C. briggsae* exceeds one substitution per site (Stein et al. 2003). This situation was only very recently improved with the discovery and genome sequencing of *Caenorhabditis inopinata,* a species that is estimated to have diverged from *C. elegans* only ~10 million years ago (Kanzaki et al. 2018).

As outlined in Chapter 3, multispecies whole-genome alignments (WGAs) are powerful resources for identifying novel coding sequences and conserved noncoding elements. The power to distinguish between coding and noncoding sequence, or conserved and nonconserved sequence, is provided by the neutral branch length connecting the species included in the WGA. While total branch length can be increased by including distantly related species, the effect of this is often negligible, since the genomes of these species cannot be sufficiently aligned. Thus, an optimal WGA will maximise neutral branch length and alignability by including a large number of species that are relatively closely related (i.e. <<1 substitution on average per neutrally evolving site (Hiller et al. 2013)). Substantial sequencing effort has been applied in order to produce such datasets for model species, with examples including the 12 species *Drosophila* WGA (Stark et al. 2007), 29 mammals WGA (Lindblad-Toh et al. 2011) and nine species Brassicaceae WGA (Haudry et al. 2013). With the rapid increase in the number of sequenced genomes and subsequent improvements in alignment algorithms, WGAs are now growing to enormous sizes, most notably a 240 species mammalian WGA (Zoonomia Consortium 2020) and a 605 species amniote WGA (Armstrong et al. 2019), providing the power to identify conserved elements at single base resolution. Although such resolution is not currently an attainable goal for *Chlamydomonas*, with only three closely related species currently available for alignment, it would be highly desirable to increase this number towards that achieved for other model organisms in the past decade or so (i.e. by having ten or more species). This would be of broad interest to the *Chlamydomonas* community, since it would enable general annotation improvements in both coding sequences and regulatory elements. It would also substantially increase our understanding of how selective constraint varies across the *C. reinhardtii* genome and would allow us to analyse and contrast patterns of selection acting on coding and functional noncoding sites. As with the shortfalls in population genetics datasets, remedying this situation is a question of sampling: how can we reliably isolate *C. reinhardtii* and its close relatives?

6.3 Trawling for *Chlamydomonas*

With the rapid improvements in sequencing technology and the associated fall in costs (1.4.1, 1.4.2), sequencing effort is no longer a major barrier to producing large scale population genetics and comparative genomics datasets, especially for species with moderately sized genomes. The work in this thesis utilises sequences from almost all known isolates of *C. reinhardtii* and its close relatives, and much of the potential to further develop *Chlamydomonas* as a study system for evolutionary biology is dependent on sampling additional isolates and species. Sampling of *C. reinhardtii* has previously been a laborious and challenging process. Soil samples would be collected and cultured, and candidate individuals would be identified morphologically and tested by mating with laboratory strains (Gross et al. 1988; Spanier et al. 1992; Sack et al. 1994). Attempts using this approach generally sampled several locations with very low success rates, suggesting that *C. reinhardtii* is not particularly abundant, or alternatively that the difficulty of morphological identification resulted in many potential isolates being missed. Furthermore, by relying on mating tests, other close relatives that may be biologically significant are likely to have been overlooked. Thus, developing a high-throughput screen for the identification of *C. reinhardtii* and related species is paramount.

In principle, it is not difficult to obtain unialgal cultures of motile photosynthetic algae in a random and high-throughput manner. Soil samples can be collected at any time of year and either dried or frozen to reduce the presence of contaminating species. After adding culture media to a soil sample and incubating for a short period of time, zygospores will germinate and vegetative algal cultures will begin to grow. The algae in these cultures can then be screened for phototaxis to separate nonmotile and motile species (Sack et al. 1994), before being diluted and plated on agar. At the correct level of dilution, separate colonies will grow from single individuals, which can then be picked and transferred to individual liquid cultures. In the past, each resulting culture would be morphologically assessed, and a mating test would be performed if a culture was deemed sufficiently similar to *C. reinhardtii*. Now that several genomic sequences are available across the core-*Reinhardtinia*, it should be possible to design species and clade-specific PCR primers. For example, a set of three primer pairs could potentially be designed that respectively amplified solely in *C. reinhardtii*, in *Chlamydomonas* but not *Edaphochlamys*, and in *Chlamydomonas* and *Edaphochlamys* but not volvocine algae. Colony PCR, which bypasses the requirement to perform individual DNA extractions, has been developed for *C. reinhardtii* (Cao et al. 2009; Wan et al. 2011) and has recently been applied in 96-well plates on thousands cultures (Nouemssi et al. 2020). For each soil sample, 96 colonies could be randomly chosen and transferred to the wells of a 96 well plate. After a short period of growth, a small aliquot from each well would be transferred to a 96 well PCR plate, PCR would be performed, and amplification scored by gel electrophoresis. Cultures that amplified successfully would then be immediately available from the original 96 well plate, from which they could be transferred to more permanent culture stocks. It may even be possible to perform a preliminary PCR on the initial multi-algal cultures to first identify if any species of interest are present in the soil sample at all. Although such an experimental design is currently entirely hypothetical, if it could be implemented, we may be able to improve our understanding of the environments in which *Chlamydomonas* species are abundant. This would potentially enable more targeted future sampling, hopefully increasing sampling efficiency. Developing such a high-throughput screen would be a major breakthrough in *Chlamydomonas* research and will be priority of future work.

6.4 Concluding remarks

C. reinhardtii is a promising model system to study fundamental evolutionary processes, especially via the combination of experimental results with inferences from population genetics and comparative genomics analyses. The species also has potential to be developed as a model to study the evolutionary ecology of microbial eukaryotes from a genomic perspective, a topic that has been severely neglected. In this thesis, I have taken some of the first steps towards developing *C. reinhardtii* and *Chlamydomonas* in these directions. In the process, I have produced a number of datasets and resources that will be useful for future evolutionary analyses in the species, and should also find extensive use in the wider *Chlamydomonas* community. However, to fully realise the potential of *C. reinhardtii* as an evolutionary model, we must increase our ecological knowledge of the species and develop reliable sampling methods. I hope to see this realised in the near future, and I greatly look forward to observing the continued development of *Chlamydomonas* as a study system for evolutionary genomics research.

References

Abascal F, Juan D, Jungreis I, Kellis M, Martinez L, Rigau M, Rodriguez JM, Vazquez J, Tress ML. 2018. Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res* **46**: 7070-7084.

Adams M, McBroome J, Maurer N, Pepper-Tunick E, Saremi NF, Green RE, Vollmers C, Corbett-Detig RB. 2020. One fly-one genome: chromosome-scale genome assembly of a single outbred *Drosophila melanogaster*. *Nucleic Acids Res* **48**: e75.

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. *Genome Biol Evol* **9**: 1329-1340.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.

Alföldi J, Di Palma F, Grabherr M, Williams C, Kong LS, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**: 587-591.

Alföldi J, Lindblad-Toh K. 2013. Comparative genomics as a tool to understand evolution and disease. *Genome Res* **23**: 1063-1068.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* **17**: 81-92.

Aoyama H, Kuroiwa T, Nakamura S. 2008. Observations of chromosomal behaviour in living meiotic zygotes of Chlamydomonas reinhardtii (Chlorophyceae). *Eur J Phycol* **43**: 389-394.

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.

Aravind L, Koonin EV. 2000. SAP - a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem Sci* **25**: 112-114.

Aravind L, Walker DR, Koonin EV. 1999. Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res* **27**: 1223-1242.

Arkhipova IR. 2006. Distribution and phylogeny of Penelope-like elements in eukaryotes. *Syst Biol* **55**: 875-885.

Arkhipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mobile DNA* **8**.

Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. 2003. Retroelements containing introns in diverse invertebrate taxa. *Nat Genet* **33**: 123-124.

Arkhipova IR, Yushenova IA. 2019. Giant transposons in eukaryotes: is bigger better? *Genome Biol Evol* **11**: 906-918.

Arkhipova IR, Yushenova IA, Rodriguez F. 2013. Endonuclease-containing *Penelope* retrotransposons in the bdelloid rotifer *Adineta vaga* exhibit unusual structural features and play a role in expansion of host gene families. *Mob DNA* **4**: 19.

Arkhipova IR, Yushenova IA, Rodriguez F. 2017. Giant reverse transcriptase-encoding transposable elements at telomeres. *Mol Biol Evol* **34**: 2245-2257.

Armstrong J, Hickey G, Diekhans M, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J et al. 2019. Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. *bioRxiv*.

Baas Becking LGM. 1934. *Geobiologie of Inleiding tot de Milieukunde*. Van Stockum & Zoon, The Hague.

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.

Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice W, Valenzuela N. 2011. Are all sex chromosomes created equal? *Trends Genet* **27**: 350-357.

Baclaocos J, Santesmasses D, Mariotti M, Bierla K, Vetick MB, Lynch S, McAllen R, Mackrill JJ, Loughran G, Guigo R et al. 2019. Processive recoding and metazoan evolution of selenoprotein P: up to 132 UGAs in molluscs. *J Mol Biol* **431**: 4381-4407.

Baier T, Wichmann J, Kruse O, Lauersen KJ. 2018. Intron-containing algal transgenes mediate efficient recombinant gene expression in the green microalga *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **46**: 6909-6919.

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.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol* **5**: e310.

Bensch S, Akesson M. 2005. Ten years of AFLP in ecology and evolution: why so few animals? *Mol Ecol* **14**: 2899-2914.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Bentley DR. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**: 545-552.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

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**.

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.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Böndel KB, Kraemer SA, Samuels T, McClean D, Lachapelle J, Ness RW, Colegrave N, Keightley PD. 2019. Inferring the distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii*. *PLoS Biol* **17**: e3000192.

Booker TR, Jackson BC, Keightley PD. 2017. Detecting positive selection in the genome. *BMC Biol* **15**: 98.

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.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**.

Boynton JE, Harris EH, Burkhart BD, Lamerson PM, Gillham NW. 1987. Transmission of mitochondrial and chloroplast genomes in crosses of *Chlamydomonas*. *Proc Natl Acad Sci U S A* **84**: 2391-2395.

Brand H, Collins RL, Hanscom C, Rosenfeld JA, Pillalamarri V, Stone MR, Kelley F, Mason T, Margolin L, Eggert S et al. 2015. Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am J Hum Genet* **97**: 170-176.

Bryszewska MA, Måge A. 2015. Determination of selenium and its compounds in marine organisms. *J Trace Elem Med Biol* **29**: 91-98.

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

Bush SJ. 2020. Read trimming has minimal effect on bacterial SNP-calling accuracy. *Microb Genom* **6**.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012-2018.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cao M, Fu Y, Guo Y, Pan J. 2009. *Chlamydomonas* (Chlorophyceae) colony PCR. *Protoplasma* **235**: 107-110.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Carmi S, Palamara PF, Vacic V, Lencz T, Darvasi A, Pe'er I. 2013. The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics* **193**: 911-928.

Carmi S, Wilton PR, Wakeley J, Pe'er I. 2014. A renewal theory approach to IBD sharing. *Theoretical Population Biology* **97**: 35-48.

Caron DA. 2009. Past President's address: protistan biogeography: why all the fuss? *J Eukaryot Microbiol* **56**: 105-112.

Carriconde F, Gardes M, Jargeat P, Heilmann-Clausen J, Mouhamadou B, Gryta H. 2008. Population evidence of cryptic species and geographical structure in the cosmopolitan ectomycorrhizal fungus, *Tricholoma scalpturatum*. *Microbial Ecology* **56**: 513-524.

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.

Castanera R, Perez G, Lopez L, Sancho R, Santoyo F, Alfaro M, Gabaldon T, Pisabarro AG, Oguiza JA, Ramirez L. 2014. Highly expressed captured genes and cross-kingdom domains present in Helitrons create novel diversity in *Pleurotus ostreatus* and other fungi. *Bmc Genomics* **15**: 1071.

Casteleyn G, Leliaert F, Backeljau T, Debeer AE, Kotaki Y, Rhodes L, Lundholm N, Sabbe K, Vyverman W. 2010. Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proc Natl Acad Sci U S A* **107**: 12952-12957.

Castellano S, Novoselov SV, Kryukov GV, Lescure A, Blanco E, Krol A, Gladyshev VN, Guigo R. 2004. Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep* **5**: 71-77.

Castruita M, Casero D, Karpowicz SJ, Kropat J, Vieler A, Hsieh SI, Yan W, Cokus S, Loo JA, Benning C et al. 2011. Systems biology approach in *Chlamydomonas* reveals connections between copper nutrition and multiple metabolic steps. *Plant Cell* **23**: 1273-1292.

Cervera A, De la Peña M. 2014. Eukaryotic *Penelope*-like retroelements encode hammerhead ribozyme motifs. *Mol Biol Evol* **31**: 2941-2947.

Cervera A, de la Peña M. 2020. Small circRNAs with self-cleaving ribozymes are highly expressed in diverse metazoan transcriptomes. *Nucleic Acids Res* **48**: 5054-5064.

Cervera A, Urbina D, de la Peña M. 2016. Retrozymes are a unique family of non-autonomous retrotransposons with hammerhead ribozymes that propagate in plants through circular RNAs. *Genome Biol* **17**: 135.

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**.

Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **18**: 495-506.

Charlesworth B. 1978. The population genetics of anisogamy. *J Theor Biol* **73**: 347-357.

Charron G, Leducq JB, Landry CR. 2014. Chromosomal variation segregates within incipient species and correlates with reproductive isolation. *Mol Ecol* **23**: 4362-4372.

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* **49**: 7571-7587.

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.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J* **89**: 789-804.

Cheng H, Concepcion T, Feng X, Zhang H, Li H. 2020. Haplotype-resolved de novo assembly with phased assembly graphs. *arXiv* doi:arXiv:2008.01237

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**: 1050-1054.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71-86.

Clement M, Snell Q, Walker P. 2002. TCS: Estimating gene genealogies. *Proceedings of the 16th International Parallel and Distributed Processing Symposium* **2:184**.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71-76.

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.

Colegrave N. 2002. Sex releases the speed limit on evolution. *Nature* **420**: 664-666.

Colegrave N, Kaltz O, Bell G. 2002. The ecology and genetics of fitness in *Chlamydomonas*. VIII. The dynamics of adaptation to novel environments after a single episode of sex. *Evolution* **56**: 14-21.

Coleman AW, Mai JC. 1997. Ribosomal DNA ITS-1 and ITS-2 sequence comparisons as a tool for predicting genetic relatedness. *J Mol Evol* **45**: 168-177.

Collins S, Bell G. 2004. Phenotypic consequences of 1,000 generations of selection at elevated CO2 in a green alga. *Nature* **431**: 566-569.

Collins S, Sultemeyer D, Bell G. 2006. Changes in C uptake in populations of *Chlamydomonas reinhardtii* selected at high CO2. *Plant Cell Environ* **29**: 1812-1819.

Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**: 987-991.

Cosby RL, Chang NC, Feschotte C. 2019. Host-transposon interactions: conflict, cooperation, and cooption. *Genes Dev* **33**: 1098-1116.

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, 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* **In press**.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188-1190.

Crosby MA, Gramates LS, Dos Santos G, Matthews BB, St Pierre SE, Zhou P, Schroeder AJ, Falls K, Emmert DB, Russo SM et al. 2015. Gene model annotations for *Drosophila melanogaster*: the rule-benders. *G3 (Bethesda)* **5**: 1737-1749.

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**.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Dangeard PA. 1888. Recherches sur les algues inférieures. *Annales des Sciences Naturelles, Botanique, série 7*: 105-175.

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**.

de la Peña M, Ceprián R, Cervera A. 2020. A singular and widespread group of mobile genetic elements: RNA circles with autocatalytic ribozymes. *Cells* **9**.

De Meester L, Gómez A, Okamura B, Schwenk K. 2002. The Monopolization Hypothesis and the dispersal-gene flow paradox in aquatic organisms. *Acta Oecologica* **23**: 121-135.

Del Cortona A, Jackson CJ, Bucchini F, Van Bel M, D'hondt S, Skaloud P, Delwiche CF, Knoll AH, Raven JA, Verbruggen H et al. 2020. Neoproterozoic origin and multiple transitions to macroscopic growth in green seaweeds. *Proc Natl Acad Sci U S A* **117**: 2551-2559.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.

Ding J, Li X, Hu H. 2012. Systematic prediction of cis-regulatory elements in the *Chlamydomonas reinhardtii* genome using comparative genomics. *Plant Physiol* **160**: 613-623.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.

Douglas TE, Kronforst MR, Queller DC, Strassmann JE. 2011. Genetic diversity in the social amoeba *Dictyostelium discoideum*: population differentiation and cryptic species. *Molecular Phylogenetics and Evolution* **60**: 455-462.

Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **9**: e92621.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203-218.

Duncan L, Bouckaert K, Yeh F, Kirk DL. 2002. kangaroo, a mobile element from Volvox carteri, is a member of a newly recognized third class of retrotransposons. *Genetics* **162**: 1617-1630.

Dunin-Horkawicz S, Feder M, Bujnicki JM. 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *Bmc Genomics* **7**: 98.

Dupeyron M, Singh KS, Bass C, Hayward A. 2019. Evolution of Mutator transposable elements across eukaryotic diversity. *Mob DNA* **10**: 12.

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

Dürrenberger F, Thompson AJ, Herrin DL, Rochaix JD. 1996. Double strand break-induced recombination in *Chlamydomonas reinhardtii* chloroplasts. *Nucleic Acids Res* **24**: 3323-3331.

Dutcher SK. 2014. The awesome power of dikaryons for studying flagella and basal bodies in *Chlamydomonas reinhardtii*. *Cytoskeleton* **71**: 79-94.

Dutcher SK, Power J, Galloway RE, Porter ME. 1991. Reappraisal of the genetic map of *Chlamydomonas reinhardtii*. *J Hered* **82**: 295-301.

Earl D, Nguyen N, Hickey G, Harris RS, Fitzgerald S, Beal K, Seledtsov I, Molodtsov V, Raney BJ, Clawson H et al. 2014. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res* **24**: 2077-2089.

Earl DA, Vonholdt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**: 359-361.

Eberlein C, Henault M, Fijarczyk A, Charron G, Bouvier M, Kohn LM, Anderson JB, Landry CR. 2019. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat Commun* **10**: 923.

Ehrenberg CG. 1838. *Die infusionsthierchen als vollkommene organismen: Ein blick in das tiefere organische leben der natur*. L. Voss, Leipzig, Germany.

Eickbush TH. 2002. R2 and related site-specific non-long terminal repeat retrotransposons. In *Mobile DNA*, Vol II (ed. N Craig, et al.), pp. 813-835. ASM Press, Washington, DC.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J et al. 2020. Pangenome graphs. *Annu Rev Genomics Hum Genet* **21**: 139-162.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc Natl Acad Sci U S A* **108**: 2831-2836.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**: 157.

Ettl H. 1976. Die gattung *Chlamydomonas* Ehrenberg (*Chlamydomonas* und die nächstverwandten gattungen II). *Beih Nova Hedwigia* **49**: 1-1122.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611-2620.

Eversole RA. 1956. Biochemical mutants of *Chlamydomonas reinhardi*. *Am J Bot* **43**: 404-407.

Evgen'ev MB, Arkhipova IR. 2005. Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res* **110**: 510-521.

Evgen'ev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, Lankenau DH, Corces VG. 1997. *Penelope*, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. *Proc Natl Acad Sci U S A* **94**: 196-201.

Faircloth BC, Branstetter MG, White ND, Brady SG. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Resour* **15**: 489-501.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol* **61**: 717-726.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567-1587.

Fang Y, Coelho MA, Shu H, Schotanus K, Thimmappa BC, Yadav V, Chen H, Malc EP, Wang J, Mieczkowski PA et al. 2020. Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in Stramenopila-Alveolata-Rhizaria lineages. *PLoS Genet* **16**: e1008646.

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

Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE, Olson B, Nozaki H, Durand PM. 2018. The 4-Celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocine lineage. *Mol Biol Evol* **35**: 855-870.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Fenchel T, Finlay BJ. 2004. The ubiquity of small species: Patterns of local and global diversity. *Bioscience* **54**: 777-784.

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.

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, Armbrust EV, Goodenough UW. 2002. Genetic structure of the mating-type locus of *Chlamydomonas reinhardtii*. *Genetics* **160**: 181-200.

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

Ferris PJ, Pavlovic C, Fabry S, Goodenough UW. 1997. Rapid evolution of sex-related genes in Chlamydomonas. *Proc Natl Acad Sci U S A* **94**: 8634-8639.

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.

Feschotte C, Pritham EJ. 2005. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet* **21**: 551-552.

Filatov DA. 2019. Extreme Lewontin's paradox in ubiquitous marine phytoplankton species. *Mol Biol Evol* **36**: 4-14.

Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science* **296**: 1061-1063.

Finlay BJ, Fenchel T. 1999. Divergent perspectives on protist species richness. *Protist* **150**: 229-233.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet* **5**: 103-107.

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.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461-465.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**: 9451-9457.

Foissner W. 1999. Protist diversity: estimates of the near-imponderable. *Protist* **150**: 363-368.

Foissner W. 2006. Biogeography and dispersal of micro-organisms: A review emphasizing protists. *Acta Protozool* **45**: 111-136.

Foissner W. 2008. Protist diversity and distribution: some basic considerations. *Biodivers Conserv* **17**: 235-242.

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**: 3702-3704.

Fryxell GA. 1983. *Survival strategies of the algae*. Cambridge University Press, New York, NY.

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.

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.

Furuyama S, Biggins S. 2007. Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proc Natl Acad Sci U S A* **104**: 14706-14711.

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.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv* doi:arXiv:1207.3907.

Gel B, Serra E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**: 3088-3090.

Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**: 1775-1787.

Gladyshev EA, Arkhipova IR. 2007. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A* **104**: 9352-9357.

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

Goodenough UW. 1992. Green yeast. *Cell* **70**: 533-538.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333-351.

Goodwin TJ, Butler MI, Poulter RT. 2003. Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology* **149**: 3099-3109.

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.

Gordon SP, Tseng E, Salamov A, Zhang JW, Meng XD, Zhao ZY, Kang DW, Underwood J, Grigoriev IV, Figueroa M et al. 2015. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *Plos One* **10**.

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

Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutierrez S, Castoe TA et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**: 1254449.

Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, Dusheyko S, Nikitin R, Mondo SJ, Salamov A et al. 2021. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res* **49**: D1004-D1011.

Gross CH, Ranum LP, Lefebvre PA. 1988. Extensive restriction fragment length polymorphisms in a new isolate of *Chlamydomonas reinhardtii*. *Current Genetics* **13**: 503-508.

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. *Eukaryot Cell* **2**: 1137-1150.

Guerreiro MPG. 2012. What makes transposable elements move in the *Drosophila* genome? *Heredity* **108**: 461-468.

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

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654-5666.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res* **16**: 875-884.

Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu Rev Ecol Evol S* **40**: 151-172.

Hamaji T, Ferris PJ, Nishii I, Nishimura Y, Nozaki H. 2013. Distribution of the sex-determining gene *MID* and molecular correspondence of mating types within the isogamous genus *Gonium* (Volvocales, Chlorophyta). *PLoS One* **8**: e64385.

Hamaji T, Kawai-Toyooka H, Toyoda A, Minakuchi Y, Suzuki M, Fujiyama A, Nozaki H, Smith DR. 2017. Multiple independent changes in mitochondrial genome conformation in Chlamydomonadalean algae. *Genome Biol Evol* **9**: 993-999.

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**.

Hamaji T, Lopez D, Pellegrini M, Umen J. 2016a. Identification and characterization of a *cis*-regulatory element for zygotic gene expression in *Chlamydomonas reinhardtii*. *G3 (Bethesda)* **6**: 1541-1548.

Hamaji T, Mogi Y, Ferris PJ, Mori T, Miyagishima S, Kabeya Y, Nishimura Y, Toyoda A, Noguchi H, Fujiyama A et al. 2016b. Sequence of the *Gonium pectorale* mating locus reveals a complex and dynamic history of changes in volvocine algal mating haplotypes. *G3 (Bethesda)* **6**: 1179-1189.

Hanschen ER, Marriage TN, Ferris PJ, Hamaji T, Toyoda A, Fujiyama A, Neme R, Noguchi H, Minakuchi Y, Suzuki M et al. 2016. The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nat Commun* **7**: 11370.

Hanson G, Coller J. 2018. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19**: 20-30.

Harris EH. 1989. *The Chlamydomonas Sourcebook: a comprehensive guide to biology and laboratory use*. Academic Press.

Harris EH. 2001. *Chlamydomonas* as a model organism. *Annual Review of Plant Physiology and Plant Molecular Biology* **52**: 363-406.

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

Harris EH, Boynton JE, Gillham NW, Burkhart BD, Newman SM. 1991. Chloroplast genome organization in *Chlamydomonas*. *Arch Protistenkd* **139**: 183-192.

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.

Hastings PJ, Levine EE, Cosbey E, Hudock MO, Gillham NW, Surzycki SJ, Loppes R, Levine RP. 1965. The linkage groups of *Chlamydomonas reinhardi*. *Microb Genet Bull* **B23**: 17-19.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**: 891-898.

Heger TJ, Mitchell EA, Leander BS. 2013. Holarctic phylogeography of the testate amoeba *Hyalosphenia papilio* (Amoebozoa: Arcellinida) reveals extensive genetic diversity explained more by environment than dispersal limitation. *Mol Ecol* **22**: 5172-5184.

Herron MD, Borin JM, Boswell JC, Walker J, Chen IK, Knox CA, Boyd M, Rosenzweig F, Ratcliff WC. 2019. *De novo* origins of multicellularity in response to predation. *Sci Rep* **9**: 2328.

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.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6**: e1001115.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**: 1341-1342.

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.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **6**: e1001107.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226-231.

Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, Bejerano G. 2013. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. *Nucleic Acids Res* **41**: e151.

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res* **15**: 1651-1660.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* **5**: 183-188.

Hirashima T, Tajima N, Sato N. 2016. Draft genome sequences of four species of *Chlamydomonas* containing phosphatidylcholine. *Microbiol Resour Ann* **4**.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* **35**: 518-522.

Hoeijmakers WA, Flueck C, Francoijs KJ, Smits AH, Wetzel J, Volz JC, Cowman AF, Voss T, Stunnenberg HG, Bartfai R. 2012. *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. *Cell Microbiol* **14**: 1391-1401.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**: 767-769.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. *Methods in Molecular Biology* **1962**: 65-95.

Hong JW, Jeong J, Kim SH, Kim S, Yoon HS. 2013. Isolation of a Korean domestic microalga, *Chlamydomonas reinhardtii* KNUA021, and analysis of its biotechnological potential. *Journal of Microbiology and Biotechnology* **23**: 375-381.

Hoshaw RW. 1965. Mating types of *Chlamydomonas* from the collection of Gilbert M. Smith. *Journal of Phycology* **1**: 194-&.

Hoshaw RW, Ettl H. 1966. *Chlamydomonas smithii* sp. nov. - a chlamydomonad interfertile with *Chlamydomonas reinhardtii*. *Journal of Phycology* **2**: 93-96.

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F et al. 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* **316**: 1625-1628.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R et al. 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res* **25**: 445-458.

Housman G, Ulitsky I. 2016. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim Biophys Acta* **1859**: 31-40.

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.

Hua J, Smith DR, Borza T, Lee RW. 2012. Similar relative mutation rates in the three genetic compartments of *Mesostigma* and *Chlamydomonas*. *Protist* **163**: 105-115.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583-589.

Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* **16**: 294.

Hyams J, Davies DR. 1972. Induction and characterization of cell-wall mutants of *Chlamydomonas reinhardi*. *Mutat Res* **14**: 381-&.

Immler S, Otto SP. 2015. The evolution of sex chromosomes in organisms with separate haploid sexes. *Evolution* **69**: 694-708.

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.

Jackson BC, Campos JL, Haddrill PR, Charlesworth B, Zeng K. 2017. Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in *Drosophila*. *Genome Biol Evol* **9**: 102-123.

Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. 2018a. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**: i748-i756.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et al. 2018b. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338-345.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801-1806.

Jang H, Ehrenreich IM. 2012. Genome-wide characterization of genetic variation in the unicellular, green alga *Chlamydomonas reinhardtii*. *PLoS One* **7**: e41307.

Jangam D, Feschotte C, Betran E. 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet* **33**: 817-831.

Jeong BR, Wu-Scharf D, Zhang C, Cerutti H. 2002. Suppressors of transcriptional transgenic silencing in *Chlamydomonas* are sensitive to DNA-damaging agents and reactivate transposable elements. *Proc Natl Acad Sci U S A* **99**: 1076-1081.

Jiang J, Wang Y, Susac L, Chan H, Basu R, Zhou ZH, Feigon J. 2018. Structure of Telomerase with Telomeric DNA. *Cell* **173**: 1179-1190 e1113.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun* **11**: 989.

Johri P, Krenek S, Marinov GK, Doak TG, Berendonk TU, Lynch M. 2017. Population genomics of *Paramecium* species. *Mol Biol Evol* **34**: 1194-1216.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res* **21**: 2096-2113.

Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. *Biol Direct* **6**: 44.

Kaina B. 2004. Mechanisms and consequences of methylating agent-induced SCEs and chromosomal aberrations: a long road traveled and still a far way to go. *Cytogenet Genome Res* **104**: 77-86.

Kaltz O, Bell G. 2002. The ecology and genetics of fitness in *Chlamydomonas*. XII. Repeated sexual episodes increase rates of adaptation to novel environments. *Evolution* **56**: 1743-1753.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**: 587-589.

Kanzaki N, Tsai IJ, Tanaka R, Hunt VL, Liu D, Tsuyama K, Maeda Y, Namai S, Kumagai R, Tracey A et al. 2018. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. *Nat Commun* **9**: 3216.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* **98**: 8714-8719.

Kapitonov VV, Jurka J. 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol Biol Evol* **20**: 38-46.

Kapitonov VV, Jurka J. 2004a. L1-1_CR, a family of L1-like non-LTR retrotransposons from the green algae genome. *Repbase Reports* **4**: 39.

Kapitonov VV, Jurka J. 2004b. TE2-1_CR is a family of nonautonomous transposable elements - a consensus sequence. *Repbase Reports* **4**: 136.

Kapitonov VV, Jurka J. 2006a. Gulliver, a family of autonomous hAT transposons from the green algae genome. *Repbase Reports* **6**: 227.

Kapitonov VV, Jurka J. 2006b. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* **103**: 4540-4545.

Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet* **23**: 521-529.

Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* **9**: 411-412; author reply 414.

Kapitonov VV, Tempel S, Jurka J. 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**: 207-213.

Kassen R, Bell G. 1998. Experimental evolution in *Chlamydomonas*. IV. Selection in environments that vary through time at different scales. *Heredity* **80**: 732-741.

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.

Katju V, Bergthorsson U. 2019. Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol Evol* **11**: 136-165.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Kawasaki Y, Nakada T, Tomita M. 2015. Taxonomic revision of oil-producing green algae, *Chlorococcum Oleofaciens* (Volvocales, Chlorophyceae), and its relatives. *Journal of Phycology* **51**: 1000-1016.

Keightley PD, Campos JL, Booker TR, Charlesworth B. 2016. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. *Genetics* **203**: 975-984.

Keightley PD, Jackson BC. 2018. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic Site. *Genetics* **209**: 897-906.

Khaw YS, Khong NMH, Shaharuddin NA, Yusoff FM. 2020. A simple 18S rDNA approach for the identification of cultured eukaryotic microalgae with an emphasis on primers. *J Microbiol Methods* **172**: 105890.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487-493.

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.

Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, Kubo M, Okada Y. 2019. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* **9**: 1784.

Kojima KK. 2020. Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet Syst* **94**: 233-252.

Kojima KK, Fujiwara H. 2005. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15**: 1106-1117.

Kojima KK, Jurka J. 2011. Crypton transposons: identification of new diverse families and ancient domestication events. *Mob DNA* **2**: 12.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**: 117.

Koufopanou V, Hughes J, Bell G, Burt A. 2006. The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**: 1941-1946.

Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, Baxter SM, Derbyshire V. 1999. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-*Tev*I: coincidence of computational and molecular findings. *Nucleic Acids Res* **27**: 2115-2125.

Koyama T, Kondo H, Aoki T, Hirono I. 2013. Identification of two *Penelope*-like elements with different structures and chromosome localization in kuruma shrimp genome. *Mar Biotechnol* **15**: 115-123.

Kraemer SA, Bondel KB, Ness RW, Keightley PD, Colegrave N. 2017. Fitness change in relation to mutation number in spontaneous mutation accumulation lines of Chlamydomonas reinhardtii. *Evolution* **71**: 2918-2929.

Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**: 487-495.

Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. 2017. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Mol Biol Evol* **34**: 1770-1779.

Kristiansen J. 1996. Dispersal of freshwater algae - a review. *Hydrobiologia* **336**: 151-157.

Kropat J, Gallaher SD, Urzica EI, Nakamoto SS, Strenkert D, Tottey S, Mason AZ, Merchant SS. 2015. Copper economy in *Chlamydomonas*: prioritized allocation and reallocation of copper to respiration vs. photosynthesis. *Proc Natl Acad Sci U S A* **112**: 2644-2651.

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.

Kuehne HA, Murphy HA, Francis CA, Sniegowski PD. 2007. Allopatric divergence, secondary contact and genetic isolation in wild yeast populations. *Current Biology* **17**: 407-411.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**: 1870-1874.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

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.

Labunskyy VM, Hatfield DL, Gladyshev VN. 2014. Selenoproteins: molecular pathways and physiological roles. *Physiol Rev* **94**: 739-777.

Lachapelle J, Bell G, Colegrave N. 2015. Experimental adaptation to marine conditions by a freshwater alga. *Evolution* **69**: 2662-2675.

Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The *Drosophila* genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics* **199**: 1229-1241.

Laetsch DR, Blaxter ML. 2017. KinFin: software for taxon-aware analysis of clustered protein sequences. *G3 (Bethesda)* **7**: 3349-3357.

Lagator M, Vogwill T, Mead A, Colegrave N, Neve P. 2013. Herbicide mixtures at high doses slow the evolution of resistance in experimentally evolving populations of *Chlamydomonas reinhardtii*. *New Phytol* **198**: 938-945.

Lahr DJ, Laughinghouse HDt, Oliverio AM, Gao F, Katz LA. 2014. How discordant morphological and molecular evolution among microorganisms can revise our notions of biodiversity on Earth. *Bioessays* **36**: 950-959.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet* **8**: e1002453.

Lawson DJ, van Dorp L, Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun* **9**: 3258.

Lebret K, Tesson SVM, Kritzberg ES, Tomas C, Rengefors K. 2015. Phylogeography of the freshwater raphidophyte *Gonyostomum Semen* confirms a recent expansion in Northern Europe by a single haplotype. *Journal of Phycology* **51**: 768-781.

Leducq JB, Charron G, Samani P, Dubé AK, Sylvester K, James B, Almeida P, Sampaio JP, Hittinger CT, Bell G et al. 2014. Local climatic adaptation in a widespread microorganism. *Proceedings of the Royal Society B: Biological Sciences* **281**: 20132472.

Leducq JB, Nielly-Thibault L, Charron G, Eberlein C, Verta JP, Samani P, Sylvester K, Hittinger CT, Bell G, Landry CR. 2016. Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature Microbiology* **1**: 15003.

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.

Leigh JW, Bryant D. 2015. PopART: full-feature software for haplotype network construction. *Methods Ecol Evol* **6**: 1110-1116.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O. 2012. Phylogeny and molecular evolution of the green algae. *Crit Rev Plant Sci* **31**: 1-46.

Levy SE, Myers RM. 2016. Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* **17**: 95-115.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* **20**: 1983-1992.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* doi:arXiv:1303.3997.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843-2851.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

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 X, Patena W, Fauser F, Jinkerson RE, Saroussi S, Meyer MT, Ivanova N, Robertson JM, Yue R, Zhang R et al. 2019. A genome-wide algal mutant library and functional screen identifies genes required for eukaryotic photosynthesis. *Nat Genet* **51**: 627-635.

Lin H, Cliften PF, Dutcher SK. 2018. MAPINS, a highly efficient detection method that identifies insertional mutations and complex DNA rearrangements. *Plant Physiol* **178**: 1436-1447.

Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE et al. 2007. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res* **17**: 1823-1836.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275-282.

Lin X, Faridi N, Casola C. 2016. An ancient transkingdom horizontal transfer of Penelope-like retroelements from arthropods to conifers. *Genome Biol Evol* **8**: 1252-1266.

Lindauer A, Fraser D, Bruderlein M, Schmitt R. 1993. Reverse transcriptase families and a copia-like retrotransposon, Osser, in the green alga Volvox carteri. *FEBS Lett* **319**: 261-266.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476-482.

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. 1997. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276**: 561-567.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471-476.

Liss M, Kirk DL, Beyser K, Fabry S. 1997. Intron sequences provide a tool for high-resolution phylogenetic analysis of volvocine algae. *Current Genetics* **31**: 214-227.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337-341.

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.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**: e119.

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.

L opez-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* **in press.** doi:10.1093/molbev/msab140.

Low SC, Berry MJ. 1996. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends Biochem Sci* **21**: 203-208.

Lowe CD, Martin LE, Montagnes DJS, Watts PC. 2012. A legacy of contrasting spatial genetic structure on either side of the Atlantic-Mediterranean transition zone in a marine protist. *Proc Natl Acad Sci U S A* **109**: 20998-21003.

Lue NF, Lin YC, Mian IS. 2003. A conserved telomerase motif within the catalytic domain of telomerase reverse transcriptase is specifically required for repeat addition processivity. *Mol Cell Biol* **23**: 8440-8449.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates, Sunderland.

Lyozin GT, Makarova KS, Velikodvorskaja VV, Zelentsova HS, Khechumian RR, Kidwell MG, Koonin EV, Evgen'ev MB. 2001. The structure and evolution of *Penelope* in the *virilis* species group of *Drosophila*: an ancient lineage of retroelements. *J Mol Evol* **52**: 445-456.

Machida H, Arai F. 2003. *Atlas of tephra in and around Japan*. University of Tokyo Press (in Japanese), Tokyo.

Macmanes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics* **5**: 13.

Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 246.

Malasarn D, Kropat J, Hsieh SI, Finazzi G, Casero D, Loo JA, Pellegrini M, Wollman FA, Merchant SS. 2013. Zinc deficiency impacts CO2 assimilation and disrupts copper homeostasis in *Chlamydomonas reinhardtii*. *J Biol Chem* **288**: 10672-10683.

Mandrioli M, Manicardi GC. 2020. Holocentric chromosomes. *PLoS Genet* **16**: e1008918.

Marco Y, Rochaix JD. 1980. Organization of the nuclear ribosomal DNA of *Chlamydomonas reinhardii*. *Mol Gen Genet* **177**: 715-723.

Margulies EH, Chen CW, Green ED. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet* **22**: 187-193.

Mariotti M, Lobanov AV, Guigo R, Gladyshev VN. 2013. SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res* **41**: e149.

Matthews BB, Dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, Zhou P, Schroeder AJ, Falls K, Strelets V et al. 2015. Gene model annotations for *Drosophila melanogaster*: impact of high-throughput data. *G3 (Bethesda)* **5**: 1721-1736.

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell* **14**: 2659-2679.

McCarthy SS, Kobayashi MC, Niyogi KK. 2004. White mutants of *Chlamydomonas reinhardtii* are defective in phytoene synthase. *Genetics* **168**: 1249-1257.

McClintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21**: 197-216.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652-654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

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.

Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in *Drosophila*. *Mob DNA* **11**: 23.

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG et al. 1997. Overview of the yeast genome. *Nature* **387**: 7-65.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79-84.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly contiguous genome assemblies of 15 *Drosophila* species generated using Nanopore sequencing. *G3 (Bethesda)* **8**: 3131-3141.

Miller SM, Schmitt R, Kirk DL. 1993. Jordan, an active Volvox transposable element similar to higher plant transposons. *Plant Cell* **5**: 1125-1138.

Misumi O, Suzuki L, Nishimura Y, Sakai A, Kawano S, Kuroiwa H, Kuroiwa T. 1999. Isolation and phenotypic characterization of *Chlamydomonas reinhardtii* mutants defective in chloroplast DNA segregation. *Protoplasma* **209**: 273-282.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24**: 319-324.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Mudge JM, Jungreis I, Hunt T, Gonzalez JM, Wright JC, Kay M, Davidson C, Fitzgerald S, Seal R, Tweedie S et al. 2019. Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res* **29**: 2073-2087.

Mugal CF, Kutschera VE, Botero-Castro F, Wolf JBW, Kaj I. 2020. Polymorphism data assist estimation of the nonsynonymous over synonymous fixation rate ratio omega for closely related species. *Mol Biol Evol* **37**: 260-279.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res* **106**: 2-9.

Müller-McNicoll M, Rossbach O, Hui J, Medenbach J. 2019. Auto-regulatory feedback by RNA-binding proteins. *J Mol Cell Biol* **11**: 930-939.

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, Misawa K, Nozaki H. 2008. Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. *Molecular Phylogenetics and Evolution* **48**: 281-291.

Nakada T, Shinkawa H, Ito T, Tomita M. 2010. Recharacterization of *Chlamydomonas reinhardtii* and its relatives with new isolates from Japan. *Journal of Plant Research* **123**: 67-78.

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.

Nakada T, Tsuchida Y, Tomita M. 2019. Improved taxon sampling and multigene phylogeny of unicellular chlamydomonads closely related to the colonial volvocalean lineage Tetrabaenaceae-Goniaceae-Volvocaceae (Volvocales, Chlorophyceae). *Molecular Phylogenetics and Evolution* **130**: 1-8.

Nakamura S. 2010. Paternal inheritance of mitochondria in *Chlamydomonas*. *J Plant Res* **123**: 163-170.

National Human Genome Research Institute. 2020. The cost of sequencing a human genome. doi:https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

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.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**: 5269-5273.

Nelson DR, Chaiboonchoe A, Fu W, Hazzouri KM, Huang Z, Jaiswal A, Daakour S, Mystikou A, Arnoux M, Sultana M et al. 2019. Potential for heightened sulfur-metabolic capacity in coastal subtropical microalgae. *iScience* **11**: 450-465.

Nelson DR, Hazzouri KM, Lauersen KJ, Jaiswal A, Chaiboonchoe A, Mystikou A, Fu W, Daakour S, Dohai B, Alzahmi A et al. 2020. Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. *Cell Host Microbe* **29**: 250-266.E8.

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.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D et al. 2018. The *Chara* genome: secondary complexity and implications for plant terrestrialization. *Cell* **174**: 448-464 e424.

Noinaj N, Guillier M, Barnard TJ, Buchanan SK. 2010. TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol* **64**: 43-60.

Nouemssi SB, Ghribi M, Beauchemin R, Meddeb-Mouelhi F, Germain H, Desgagne-Penix I. 2020. Rapid and efficient colony-PCR for high throughput screening of genetically transformed *Chlamydomonas reinhardtii*. *Life (Basel)* **10**.

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*. *EMBO J* **21**: 3681-3693.

Nozaki H, Ito M, Sano R, Uchida H, Watanabe MM, Takahashi H, Kuroiwa T. 1997. Phylogenetic analysis of *Yamagishiella* and *Platydorina* (Volvocaceae, Chlorophyta) based on *rbcL* gene sequences. *Journal of Phycology* **33**: 272-278.

Nozaki H, Ohta N, Takano H, Watanabe MM. 1999. Reexamination of phylogenetic relationships within the colonial Volvocales (Chlorophyta): An analysis of *atpB* and *rbcL* gene sequences. *Journal of Phycology* **35**: 104-112.

O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. 2015. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**: 2035-2037.

O'Donnell S, Chaux F, Fischer G. 2020. Highly contiguous Nanopore genome assembly of *Chlamydomonas reinhardtii* CC-1690. *Microbiol Resour Announc* **9**: e00726-20.

O'Donnell S, Fischer G. 2020. MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinformatics* **36**: 3242-3243.

O'Malley MA. 2008. 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Biological and Biomedical Sciences* **39**: 314-325.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P et al. 2017. vegan: Community Ecology Package. R package version 2.4-5. doi:https://CRAN.R-project.org/package=vegan.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, De Meester L. 2013. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Mol Ecol* **22**: 5983-5999.

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

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.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics* **91**: 1150-1150.

Parfrey LW, Lahr DJ, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci U S A* **108**: 13624-13629.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol Biol Evol* **27**: 1226-1234.

Pascher A. 1918. Über die beziehung der reduktionsteilung zur Mendelschen spaltung. *Berichte der Deutschen Botanischen Gesellschaft* **36**: 163-168.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**: 577-591.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M et al. 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol* **10**: e1001419.

Pei J, Grishin NV. 2014. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol* **1079**: 263-271.

Penton EH, Crease TJ. 2004. Evolution of the transposable element *Pokey* in the ribosomal DNA of species in the subgenus *Daphnia* (Crustacea: Cladocera). *Mol Biol Evol* **21**: 1727-1739.

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.

Peska V, Garcia S. 2020. Origin, diversity, and evolution of telomere sequences in plants. *Front Plant Sci* **11**: 117.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergstrom A, Sigwalt A, Barre B, Freel K, Llored A et al. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**: 339-344.

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.

Platt RN, II, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol* **8**: 403-410.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32-42.

Pohl M, Theissen G, Schuster S. 2012. GC content dependency of open reading frame prediction via stop codon frequencies. *Gene* **511**: 441-446.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Res* **20**: 291-300.

Popescu CE, Borza T, Bielawski JP, Lee RW. 2006. Evolutionary rates and expression level in *Chlamydomonas*. *Genetics* **172**: 1567-1576.

Popescu CE, Lee RW. 2007. Mitochondrial genome sequence evolution in *Chlamydomonas*. *Genetics* **175**: 819-826.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

Porter ME, Knott JA, Myster SH, Farlow SJ. 1996. The dynein gene family in *Chlamydomonas reinhardtii*. *Genetics* **144**: 569-585.

Poulter RT, Goodwin TJ. 2005. DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenet Genome Res* **110**: 575-588.

Poulter RTM, Butler MI. 2015. Tyrosine recombinase retrotransposons and transposons. *Microbiol Spectr* **3**: MDNA3-0036-2014.

Preuss D, Mets L. 2002. Plant centromere functions defined by tetrad analysis and artificial chromosomes. *Plant Physiol* **129**: 421-422.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.

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.

Pröschold T, Darienko T. 2018. The distinctions between *Chlamydomonas incerta* Pascher and *Chlamydomonas globosa* J.W. Snow and their taxonomic consequences. *Notulae algarum* **56**: 1-4.

Pröschold T, Darienko T, Krienitz L, Coleman AW. 2018. *Chlamydomonas schloesseri* sp nov (Chlamydophyceae, Chlorophyta) revealed by morphology, autolysin cross experiments, and multiple gene analyses. *Phytotaxa* **362**: 21-38.

Pröschold T, Harris EH, Coleman AW. 2005. Portrait of a species: *Chlamydomonas reinhardtii*. *Genetics* **170**: 1601-1610.

Pröschold T, Marin B, Schlösser UG, Melkonian M. 2001. Molecular phylogeny and taxonomic revision of *Chlamydomonas* (Chlorophyta). I. Emendation of *Chlamydomonas* Ehrenberg and *Chloromonas* Gobi, and description of *Oogamochlamys* gen. nov. and *Lobochlamys* gen. nov. *Protist* **152**: 265-300.

Pröschold T, Silva PC. 2007. (1768) Proposal to change the listed type of *Chlamydomonas* Ehrenb., nom. cons. (*Chlorophyta*). *TAXON* **56**: 595-596.

Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V. 2016. The Arabidopsis thaliana mobilome and its impact at the species level. *Elife* **5**.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

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.

Ranum LP, Thompson MD, Schloss JA, Lefebvre PA, Silflow CD. 1988. Mapping flagellar genes in *Chlamydomonas* using restriction fragment length polymorphisms. *Genetics* **120**: 109-122.

Ratcliff WC, Herron MD, Howell K, Pentz JT, Rosenzweig F, Travisano M. 2013. Experimental evolution of an alternating uni- and multicellular life cycle in Chlamydomonas reinhardtii. *Nat Commun* **4**: 2742.

Reboud X, Bell G. 1997. Experimental evolution in *Chlamydomonas*. III. Evolution of specialist and generalist types in environments that vary in space and time. *Heredity* **78**: 507-514.

Rengefors K, Kremp A, Reusch TBH, Wood AM. 2017. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. *J Plankton Res* **39**: 165-179.

Rengefors K, Logares R, Laybourn-Parry J. 2012. Polar lakes may act as ecological islands to aquatic protists. *Mol Ecol* **21**: 3200-3209.

Ribeiro YC, Robe LJ, Veluza DS, Dos Santos CMB, Lopes ALK, Krieger MA, Ludwig A. 2019. Study of VIPER and TATE in kinetoplastids and the evolution of tyrosine recombinase retrotransposons. *Mob DNA* **10**: 34.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* **15**: 1-18.

Riddle NC, Elgin SCR. 2018. The *Drosophila* dot chromosome: where genes flourish amidst repeats. *Genetics* **210**: 757-772.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* **29**: 24-26.

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.

Rochaix JD. 1995. *Chlamydomonas reinhardtii* as the photosynthetic yeast. *Annu Rev Genet* **29**: 209-230.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Mol Biol Evol* **31**: 1750-1766.

Rose AB. 2018. Introns as gene regulators: a brick on the accelerator. *Front Genet* **9**: 672.

Roth MS, Cokus SJ, Gallaher SD, Walter A, Lopez D, Erickson E, Endelman B, Westcott D, Larabell CA, Merchant SS et al. 2017. Chromosome-level genome assembly and transcriptome of the green alga *Chromochloris zofingiensis* illuminates astaxanthin production. *Proc Natl Acad Sci U S A* **114**: E4296-E4305.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**: 155-158.

Rymarquis LA, Handley JM, Thomas M, Stern DB. 2005. Beyond complementation. Map-based cloning in *Chlamydomonas reinhardtii*. *Plant Physiol* **137**: 557-566.

Sack L, Zeyl C, Bell G, Sharbel T, Reboud X, Bernhardt T, Koelewyn H. 1994. Isolation of four new strains of *Chlamydomonas reinhardtii* (Chlorophyta) from soil samples. *Journal of Phycology* **30**: 770-773.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* **10**: 516-522.

Salomé PA, Merchant SS. 2019. A Series of fortunate events: Introducing *Chlamydomonas* as a reference organism. *Plant Cell* **31**: 1682-1707.

Sanchez R, Zhou MM. 2011. The PHD finger: a versatile epigenome reader. *Trends Biochem Sci* **36**: 364-372.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* **265**: 687-695.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-5467.

Sasso S, Stibor H, Mittag M, Grossman AR. 2018. From molecular manipulation of domesticated *Chlamydomonas reinhardtii* to survival in nature. *eLife* **7**.

Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. 2018. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal* **17**: 196.

Schlötterer C, Pemberton J. 1998. The use of microsatellites for genetic analysis of natural populations — a critical review. In *Molecular Approaches to Ecology and Evolution*, doi:https://doi.org/10.1007/978-3-0348-8948-3_4 (ed. R DeSalle, B Schierwater). Birkhäuser, Basel.

Schnell RA, Lefebvre PA. 1993. Isolation of the *Chlamydomonas* regulatory gene NIT2 by transposon tagging. *Genetics* **134**: 737-747.

Schostak N, Pyatkov K, Zelentsova E, Arkhipova I, Shagin D, Shagina I, Mudrik E, Blintsov A, Clark I, Finnegan DJ et al. 2008. Molecular dissection of Penelope transposable element regulatory machinery. *Nucleic Acids Res* **36**: 2522-2529.

Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, Antoniou P, Knight SJL, Camps C, Pentony MM et al. 2020. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet Med* **22**: 85-94.

Scranton MA, Ostrand JT, Fields FJ, Mayfield SP. 2015. *Chlamydomonas* as a model for biofuels and bio-products production. *The Plant Journal* **82**: 523-531.

Sfeir A, Symington LS. 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem Sci* **40**: 701-714.

Sharp PM, Li WH. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281-1295.

Shoemaker WR, Lennon JT. 2018. Evolution with a seed bank: The population genetic consequences of microbial dormancy. *Evolutionary Applications* **11**: 60-75.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Sjöqvist C, Godhe A, Jonsson PR, Sundqvist L, Kremp A. 2015. Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea-Baltic Sea salinity gradient. *Mol Ecol* **24**: 2871-2885.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0. *http://wwwrepeatmaskerorg*.

Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. *http://wwwrepeatmaskerorg*.

Smith DR, Craig RJ. 2020. Does mitochondrial DNA replication in *Chlamydomonas* require a reverse transcriptase? *New Phytol* **229**: 1192-1195.

Smith DR, Lee RW. 2008. Nucleotide diversity in the mitochondrial and nuclear compartments of *Chlamydomonas reinhardtii*: investigating the origins of genome architecture. *BMC Evol Biol* **8**: 156.

Smith DR, Lee RW. 2009. Nucleotide diversity of the *Chlamydomonas reinhardtii* plastid genome: addressing the mutational-hazard hypothesis. *BMC Evol Biol* **9**: 120.

Smith DR, Lee RW. 2010. Low nucleotide diversity for the expanded organelle and nuclear genomes of Volvox carteri supports the mutational-hazard hypothesis. *Mol Biol Evol* **27**: 2244-2256.

Smith EF, Lefebvre PA. 1997. *PF20* gene product contains WD repeats and localizes to the intermicrotubule bridges in *Chlamydomonas* flagella. *Mol Biol Cell* **8**: 455-467.

Smith KM, Phatale PA, Sullivan CM, Pomraning KR, Freitag M. 2011. Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. *Mol Cell Biol* **31**: 2528-2542.

Smyth RD, Martinek GW, Ebersold WT. 1975. Linkage of six genes in *Chlamydomonas reinhardtii* and the construction of linkage test strains. *J Bacteriol* **124**: 1615-1617.

Spanier JG, Graham JE, Jarvik JW. 1992. Isolation and preliminary characterization of three *Chlamydomonas* strains interfertile with *Chlamydomonas reinhardtii* (Chlorophyta). *Journal of Phycology* **28**: 822-828.

Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205-216.

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.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**: 219-232.

Steige KA, Laenen B, Reimegard J, Scofield DG, Slotte T. 2017. Genomic analysis reveals major determinants of *cis*-regulatory variation in *Capsella grandiflora*. *Proc Natl Acad Sci U S A* **114**: 1087-1092.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* **1**: E45.

Sterken MG, Snoek LB, Kammenga JE, Andersen EC. 2015. The laboratory domestication of *Caenorhabditis elegans*. *Trends Genet* **31**: 224-231.

Stevens L, Felix MA, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frezal L, Gosse C, Kaur T et al. 2019. Comparative genomics of 10 new *Caenorhabditis* species. *Evol Lett* **3**: 217-236.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**: 2.

Storms R, Hastings PJ. 1977. A fine structure analysis of meiotic pairing in *Chlamydomonas reinhardi*. *Exp Cell Res* **104**: 39-46.

Strenkert D, Schmollinger S, Gallaher SD, Salome PA, Purvine SO, Nicora CD, Mettler-Altmann T, Soubeyrand E, Weber APM, Lipton MS et al. 2019. Multiomics resolution of molecular events during a day in the life of *Chlamydomonas*. *Proc Natl Acad Sci U S A* **116**: 2374-2383.

Suh A, Churakov G, Ramakodi MP, Platt RN, 2nd, Jurka J, Kojima KK, Caballero J, Smit AF, Vliet KA, Hoffmann FG et al. 2014. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol Evol* **7**: 205-217.

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.

Talbert PB, Henikoff S. 2020. What makes a centromere? *Experimental Cell Research* **389**.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* **320**: 486-488.

Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, Pyae Phyo A, Nosten F, Neafsey DE, Buckee CO. 2017. Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. *PLoS Genet* **13**: e1007065.

Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol* **23**: 2637-2652.

The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**: 481-491.

Thomma B, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JAL, Faino L. 2016. Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet Biol* **90**: 24-30.

Thompson EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**: 301-326.

Tulin F, Cross FR. 2016. Patching holes in the *Chlamydomonas* genome. *G3 (Bethesda)* **6**: 1899-1910.

Tzvetkov N, Breuer P. 2007. Josephin domain-containing proteins from a variety of species are active de-ubiquitination enzymes. *Biol Chem* **388**: 973-978.

Umen JG. 2020. Volvox and volvocine green algae. *Evodevo* **11**: 13.

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. 2019. Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*. *Mol Biol Evol* **36**: 1734-1745.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardti*i: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. *Curr Genet* **24**: 241-247.

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 EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The third revolution in sequencing technology. *Trends Genet* **34**: 666-681.

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.

Vanormelingen P, Evans KM, Mann DG, Lance S, Debeer AE, D'Hondt S, Verstraete T, De Meester L, Vyverman W. 2015. Genotypic diversity and differentiation among populations of two benthic freshwater diatoms as revealed by microsatellites. *Mol Ecol* **24**: 4433-4448.

Vanoverbeke J, De Meester L. 2010. Clonal erosion and genetic drift in cyclical parthenogens - the interplay between neutral and selective processes. *J Evolution Biol* **23**: 997-1012.

Vogwill T, Lagator M, Colegrave N, Neve P. 2012. The experimental evolution of herbicide resistance in Chlamydomonas reinhardtii results in a positive correlation between fitness in the presence and absence of herbicides. *J Evol Biol* **25**: 1955-1964.

Volff JN, Hornung U, Schartl M. 2001. Fish retroposons related to the *Penelope* element of *Drosophila virilis* define a new group of retrotransposable elements. *Mol Genet Genomics* **265**: 711-720.

Wakeley J, Wilton PR. 2016. Coalescent and models of identity by descent. In *Encyclopedia of Evolutionary Biology*, Vol 1 (ed. RM Kliman), pp. 287-292. Academic Press, Oxford.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Wan M, Rosenberg JN, Faruq J, Betenbaugh MJ, Xia J. 2011. An improved colony PCR procedure for genetic screening of *Chlorella* and related microalgae. *Biotechnol Lett* **33**: 1615-1619.

Wang J. 2017. The computer program STRUCTURE for assigning individuals to populations: easy to use but easier to misuse. *Mol Ecol Resour* **17**: 981-990.

Wang SC, Schnell RA, Lefebvre PA. 1998. Isolation and characterization of a new transposable element in *Chlamydomonas reinhardtii*. *Plant Molecular Biology* **38**: 681-687.

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**: 543-548.

Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* **37**: 124-126.

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.

Wheeler GL, Miranda-Saavedra D, Barton GJ. 2008. Genome analysis of the unicellular green alga *Chlamydomonas reinhardtii* indicates an ancient evolutionary origin for key pattern recognition and cell-signaling protein families. *Genetics* **179**: 193-197.

Whittaker KA, Rynearson TA. 2017. Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proc Natl Acad Sci U S A* **114**: 2651-2656.

Wiberg RA, Halligan DL, Ness RW, Necsulea A, Kaessmann H, Keightley PD. 2015. Assessing recent selection and functionality at long noncoding RNA loci in the mouse genome. *Genome Biol Evo* **7**: 2432-2444.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973-982.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet* **10**: e1004622.

Wong LH, Choo KH. 2004. Evolutionary dynamics of transposable elements at the centromere. *Trends Genet* **20**: 611-616.

Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268-272.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859-1875.

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.

Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM. 2002. A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. *Nucleic Acids Res* **30**: 2524-2537.

Wyatt MD, Pittman DL. 2006. Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks. *Chem Res Toxicol* **19**: 1580-1594.

Xia X. 2015. A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics* **199**: 573-579.

Xia X. 2018. DAMBE7: new and improved tools for data analysis in molecular biology and evolution. *Mol Biol Evol* **35**: 1550-1552.

Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. 2017. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**: 1072-1074.

Xiong W, He L, Lai J, Dooner HK, Du C. 2014. HelitronScanner uncovers a large overlooked cache of *Helitron* transposons in many plant genomes. *Proc Natl Acad Sci U S A* **111**: 10263-10268.

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, Kawai-Toyooka H, Hamaji T, Tsuchikane Y, Mori T, Takahashi F, Sekimoto H, Ferris PJ, Nozaki H. 2017a. Molecular evolutionary analysis of a gender-limited *MID* ortholog from the homothallic species *Volvox africanus* with male and monoecious spheroids. *PLoS One* **12**: e0180313.

Yamamoto R, Obbineni JM, Alford LM, Ide T, Owa M, Hwang J, Kon T, Inaba K, James N, King SM et al. 2017b. *Chlamydomonas* DYX1C1/PF23 is essential for axonemal assembly and proper morphology of inner dynein arms. *PLoS Genet* **13**: e1006996.

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.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**.

Zhang C, Wu-Scharf D, Jeong BR, Cerutti H. 2002. A WD40-repeat containing protein, similar to a fungal co-repressor, is required for transcriptional gene silencing in *Chlamydomonas*. *Plant J* **31**: 25-36.

Zhang Z, Qu C, Zhang K, He Y, Zhao X, Yang L, Zheng Z, Ma X, Wang X, Wang W et al. 2020. Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. *Curr Biol* **30**: 3330-3341 e3337.

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.

Zheng XW, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326-3328.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A* **111**: E2310-2318.

Zoonomia Consortium. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**: 240-245.

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.

Zufall RA, Dimond KL, Doerder FP. 2013. Restricted distribution and limited gene flow in the model ciliate *Tetrahymena thermophila*. *Mol Ecol* **22**: 1081-1091.

Appendix A

Supplementary Material for Chapter 1

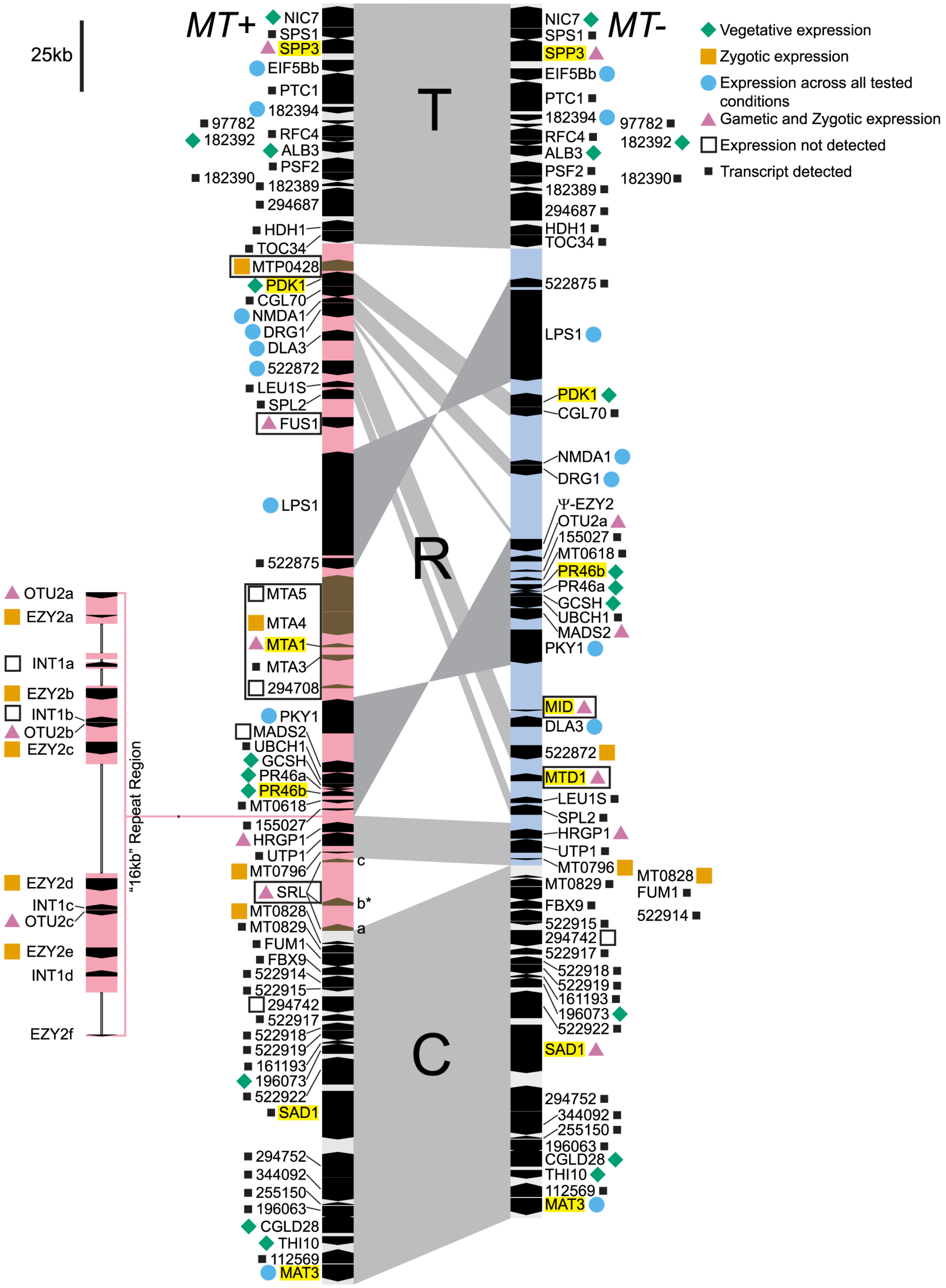


Figure S1. The *plus* and *minus* mating type loci of *C. reinhardtii*. The figure is copied and modified from De Hoff et al. (2013). The T, R (coloured pink or blue) and C domains are marked and synteny between the haplotypes is shown by grey shading. Gene orientation is shown by the direction of the triangular markers. Mating type specific genes are boxed, and genes highlighted in yellow were used for a particular purpose in the original study.

Table S1. Information on laboratory strains used in this thesis. See Gallaher et al. (2015) and the Chlamydomonas Resource Centre for further details (https://www.chlamycollection.org). See Appendix B, Table S1 for information on field isolates.

Strain	**Synonym**	**Mating Type**	**Origin**	**Use in thesis**
CC-1009	UTEX 89	–	Cambridge subline wild type	Chapter 2, population structure analyses; Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-1010	UTEX 90	+	Cambridge subline wild type	Chapter 2, population structure analyses; Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-124	137c *mt–*	–	Ebersold/Levine subline wild type	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-125	137c *mt+*	+	Ebersold/Levine subline wild type	Chapter 4, comparison to CC-503 genome and *Gypsy-7a_cRei* coverage analysis
CC-1690	21 gr	+	Sager subline wild type	Chapter 4 and 5, Nanopore genome assembly used for several purposes
CC-1691	*y1*	–	Sager subline wild type	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-503	*cw92*	+	mutagenised derivative of CC-125	Original reference genome strain, genome assembly used throughout and improved in Chapter 4
CC-4532	2137 *mt–*	–	unknown cross	Chapter 4, new reference genome strain
CC-3269	2137 *mt+*	+	CC-1690 x CC-124	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-4051	4A+	+	CC-125 x CC-124	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-407	C8	+	subclone of CC-1690, separate since 1950s	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-425	*cw15 arg2 sr-u-2-60*	+	unknown cross, no cell wall	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-4350	Matagne 302	+	unknown cross, no cell wall	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-4425	D66	+	multiple crosses, no cell wall	Chapter 4, *Gypsy-7a_cRei* coverage analysis
CC-4533	Jonikas CMJ030	–	multiple crosses	Chapter 4, *Gypsy-7a_cRei* coverage analysis

Appendix B

Supplementary Material for Chapter 2

Table S1. Summary statistics and sampling information for all genetically unique *C. reinhardtii* field isolates. Coverage depth is the mean number of reads per site, for sites with at least three mapped reads. Coverage breadth is the percentage of sites with at least three mapped reads. Genetic distances were calculated genome-wide relative to the *C. reinhardtii* reference genome (laboratory strain CC-503) using the Tamura-Nei substitution model. Table shown on following two pages.

* sampling location for CC-2344 is given as Ralston, PA, by the Chlamydomonas Resource Centre. Spanier et al. (1992) reported the location as Malverne, PA, which does not exist (there is a Malvern, PA). We have used Ralston, as there is no further information regarding the true location (Jonathon Jarvik, personal communication).
** listed as the opposite mating type in the Chlamydomonas Resource Centre.
*** listed as unknown or uncertain mating type in the Chlamydomonas Resource Centre.

Isolate	Location	Year	Sample type	Mating type	Sampling /description reference	Library	Depth of coverage (x)	Breadth of coverage (%)	Genetic distance to reference (%)	Sequencing reference
CC-1009	Amherst, MA, USA	1945	soil, potato field / laboratory	-	Hoshaw (1965)	2 x 100 bp	55.40	95.71	0.26	Gallaher et al. (2015)
CC-1010	Amherst, MA, USA	1945	” ”	+	” ”	” ”	64.98	96.32	0.06	” ”
CC-1373	South Deerfield, MA, USA	1945	soil, tobacco field	+	Hoshaw and Ettl (1966)	2 x 51 bp	84.50	96.86	0.94	Flowers et al. (2015)
CC-1952	Plymouth, MN, USA	1986	soil, lake bank	-	Gross et al. (1988)	” ”	81.05	94.85	1.97	” ”
CC-2342	Pittsburgh, PA, USA	1988	soil	-	Spanier et al. (1992)	” ”	90.81	94.13	1.88	” ”
CC-2343	Melbourne, FL, USA	1988	” ”	+	” ”	” ”	67.00	92.68	2.18	” ”
CC-2344	Ralston, PA, USA*	1988	” ”	+	” ”	” ”	87.41	95.87	1.82	” ”
CC-2931	Durham, NC, USA	1991	soil, garden	-	Pröschold et al. (2005)	” ”	78.50	95.02	1.94	” ”
CC-2932	Durham, NC, USA	1991	” ”	+	” ”	2 x 50 bp 2 x 90 bp	20.33	67.53	1.90	Jang and Ehrenreich (2012)
CC-3268	Durham, NC, USA	1991	” ”	-		2 x 150 bp	30.87	85.77	1.94	this study
CC-2935	Farnham, QC, Canada	1993	soil, arable field	-	Sack et al. (1994)	2 x 51 bp	58.56	94.64	1.24	Flowers et al. (2015)
CC-2936	Farnham, QC, Canada	1993	” ”	+	” ”	” ”	70.41	96.26	0.99	” ”
CC-2937	Farnham, QC, Canada	1993	” ”	+	” ”	” ”	90.31	96.75	1.21	” ”
CC-2938	Farnham, QC, Canada	1993	” ”	-	” ”	” ”	66.90	95.63	0.92	” ”
CC-3059	Farnham, QC, Canada	1993	” ”	- **		2 x 100 bp	29.95	92.83	1.07	Ness et al. (2016)
CC-3060	Farnham, QC, Canada	1993	” ”	+		” ”	30.82	92.09	1.10	” ”
CC-3061	Farnham, QC, Canada	1993	” ”	-		” ”	29.15	92.04	1.11	” ”
CC-3062	Farnham, QC, Canada	1993	” ”	- **		” ”	29.63	92.24	1.23	” ”
CC-3063	Farnham, QC, Canada	1993	” ”	-		” ”	29.71	92.12	1.21	” ”

Isolate	Location	Year	Sample type	Mating type	Sampling /description reference	Library	Depth of coverage (x)	Breadth of coverage (%)	Genetic distance to reference (%)	Sequencing reference
CC-3065	Farnham, QC, Canada	1993	” ”	+		” ”	30.68	92.27	1.25	” ”
CC-3068	Farnham, QC, Canada	1993	” ”	+		” ”	27.38	92.10	1.22	” ”
CC-3069	Farnham, QC, Canada	1993	” ”	+ ***		” ”	27.79	92.31	1.20	” ”
CC-3071	Farnham, QC, Canada	1993	” ”	+		” ”	28.77	92.42	1.21	” ”
CC-3073	Farnham, QC, Canada	1993	” ”	-		” ”	31.07	92.54	1.21	” ”
CC-3076	MacDonald College, QC, Canada	1994	” ”	+		” ”	28.19	92.07	1.05	” ”
CC-3079	MacDonald College, QC, Canada	1994	” ”	-		” ”	30.12	90.18	1.79	” ”
CC-3084	MacDonald College, QC, Canada	1994	” ”	-		” ”	29.93	93.10	0.89	” ”
CC-3086	MacDonald College, QC, Canada	1994	” ”	+		” ”	29.05	92.42	0.99	” ”
GB13	Farnham, QC, Canada	2016	” ”	+	this study	” ”	28.74	91.32	1.01	this study
GB66	Farnham, QC, Canada	2016	” ”	-	” ”	” ”	9.41	77.41	1.14	” ”
GB117	Farnham, QC, Canada	2016	” ”	-	” ”	” ”	21.82	88.10	0.94	” ”
GB119	Farnham, QC, Canada	2016	” ”	+	” ”	” ”	28.50	88.24	1.28	” ”
GB123	Farnham, QC, Canada	2016	” ”	-	” ”	” ”	29.70	90.13	0.99	” ”
GB138	Farnham, QC, Canada	2016	” ”	+	” ”	” ”	30.21	89.54	0.97	” ”
GB141	Farnham, QC, Canada	2016	” ”	+	” ”	” ”	30.02	88.69	1.28	” ”
NIES-2463	Kirishima, Kagoshima, Japan	2006	soil, rice paddy	+	Nakada et al. (2010)	2 x 300 bp	90.12	89.42	1.79	Arakawa *et al.* (unpublished)
NIES-2464	Kirishima, Kagoshima, Japan	2006	” ”	-	” ”	” ”	105.38	87.66	1.79	” ”

Table S2. Clonal pairs/trios of field isolates shown in table S1. The CC-1010/CC-3078 clonal pair likely represents laboratory contamination (see 2.7.3).

Included isolate	Excluded isolate(s)	Location	Year	Number of SNPs
CC-3069	CC-3064	Farnham, QC	1993	1,238
CC-1010	CC-3078	Massachusetts / MacDonald College, QC	1945/1994	245
CC-3079	CC-3075	MacDonald College, QC	1994	1,680
CC-3084	CC-3082, CC-3083	MacDonald College, QC	1994	895
GB141	GB57	Farnham, QC	2016	945

Table S3. Information for Chlamydomonas Resource Centre isolates sampled contemporaneously with the 1993 Farnham and 1994 MacDonald College *C. reinhardtii* isolates, that were shown not to be *C. reinhardtii* by sequence analysis.

Isolate	Location	*rbcL* accession number	*rbcL* best BLAST hit	Percent identity (%)
CC-3066	Farnham, QC	MN067205	*Chlamydomonas moewusii* (EF587479.1)	100%
CC-3067	” ”	/	/	/
CC-3070	” ”	MN067206	*Chlamydomonas applanata* (MK241694.1)	100%
CC-3072	” ”	MN067207	*Chlorococcum ellipsoideum* (EF113431.1)	97.2%
CC-3074	MacDonald College, QC	MN067208	*Pandorina unicocca* (D86826.1)	95.6%
CC-3077	” ”	MN067209	*Chlamydomonas applanata* (MK241694.1)	100%
CC-3080	” ”	MN067210	*Chlamydomonas applanata* (MK241694.1)	100%
CC-3081	” ”	/	/	/
CC-3085	” ”	MN067211	*Chlamydomonas peterfii* (KT624961.1)	99.7%
CC-3087	” ”	MN067212	*Chlamydomonas acidophila* (AB127987.1)	99.3%
CC-3088	” ”	MN067213	*Chlamydomonas debaryana* (MG650089.1)	99.5%
CC-3089	” ”	MN067214	*Chlamydomonas acidophila* (AB127987.1)	99.3%

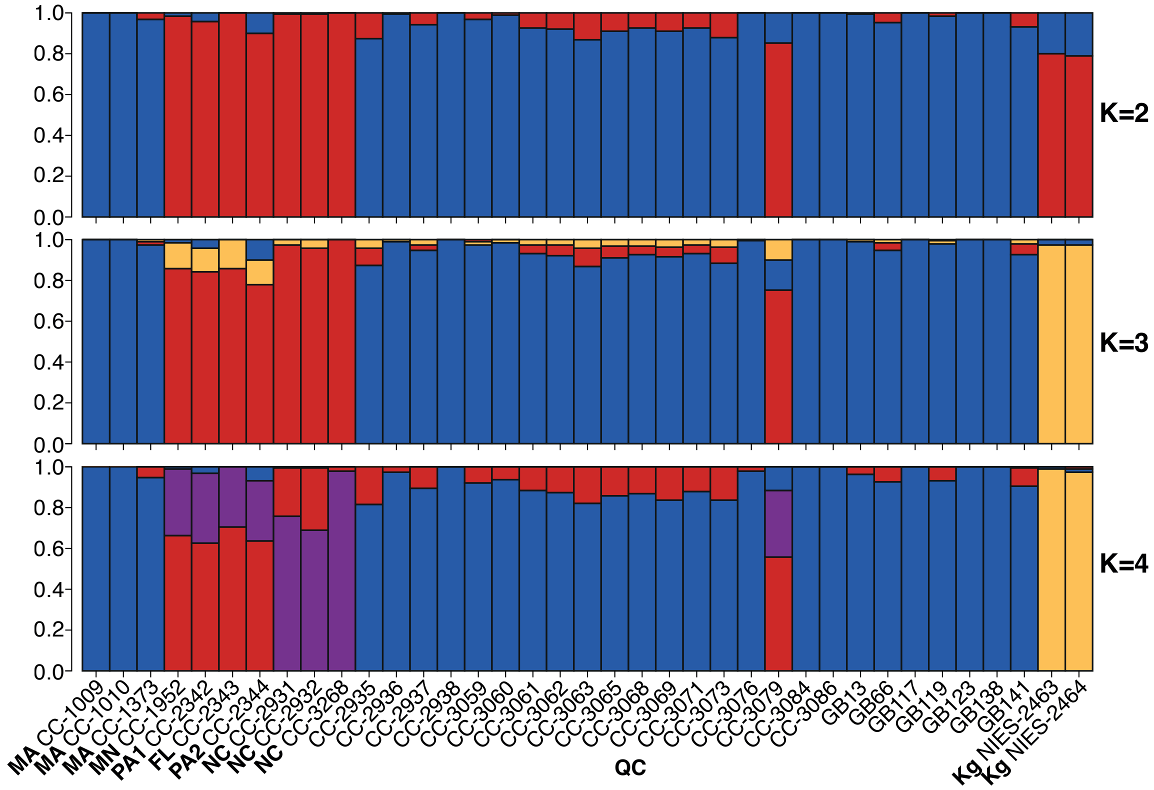


Figure S1. Species-wide STRUCTURE results shown for two, three and four ancestral populations (K).

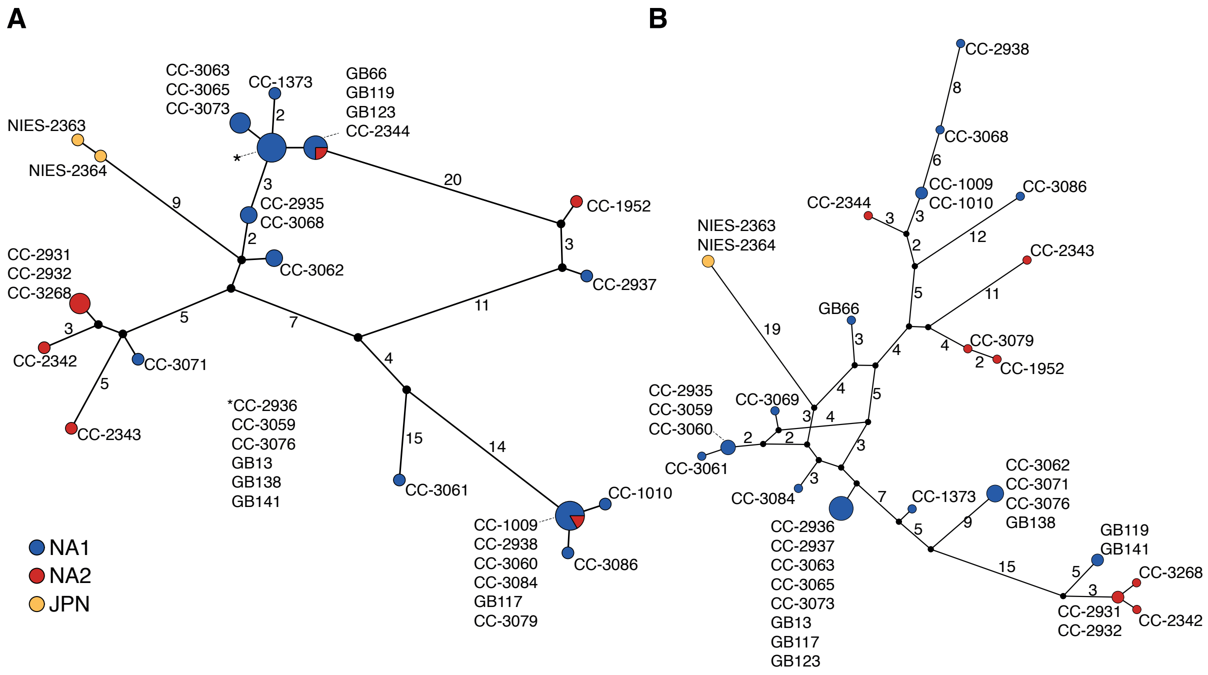


Figure S2. Organelle TCS haplotype networks, where the number on the branches (not shown to scale) represent the number of mutations between haplotypes. If no number is present, the branch represents a single mutation.

(A) Mitochondrion.

(B) Plastid coding sequence.

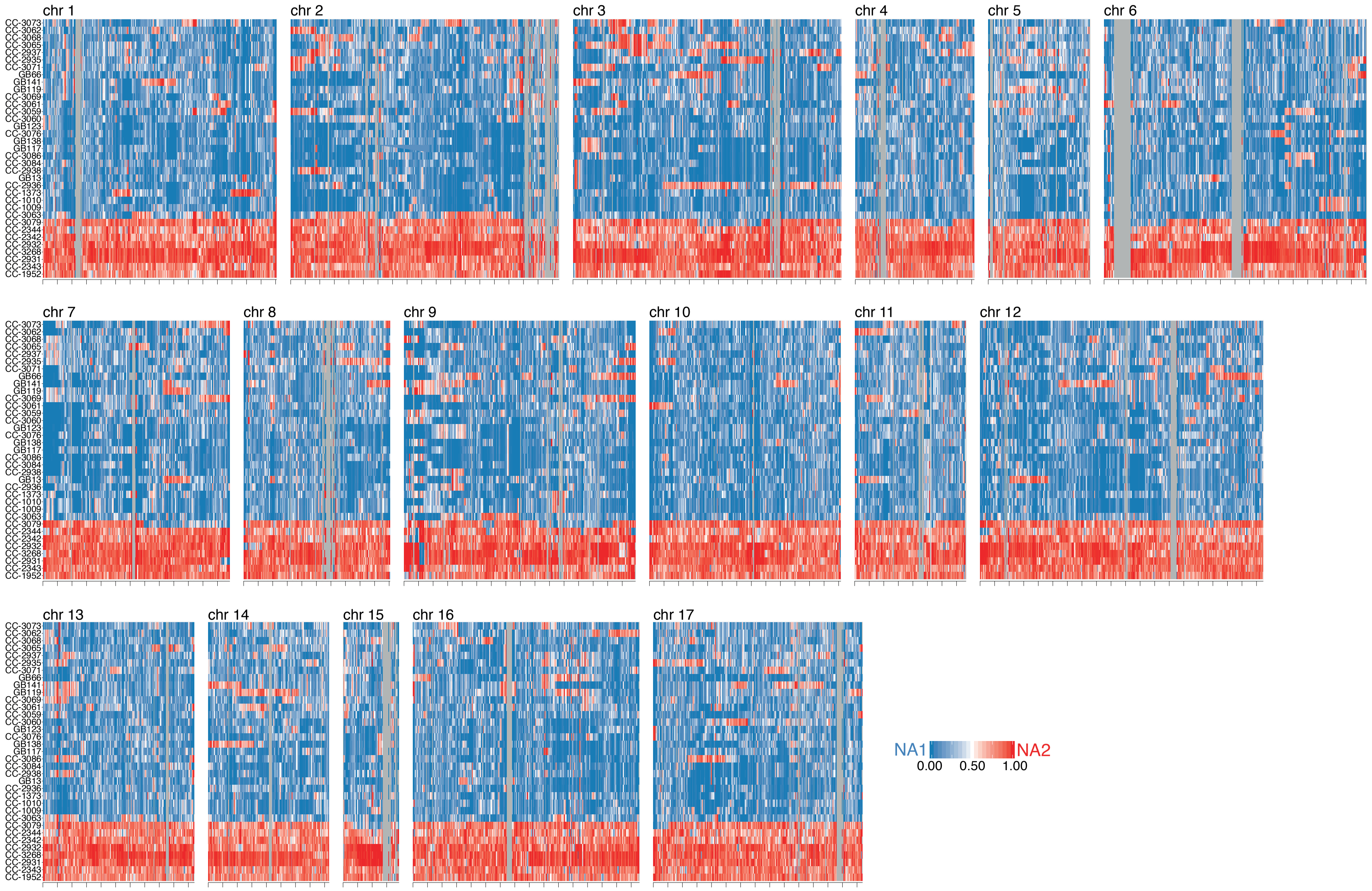


Figure S3. Admixture profiling for all chromosomes. For each isolate, the proportion of NA1 and NA2 marker SNPs in 20 kb windows along chromosomes plotted as a heat map, with 0 (dark blue) representing 100% NA1 SNPs, and 1 (dark red) representing 100% NA2 SNPs. Windows containing no sites/SNPs are shown in grey. X-axis ticks represent 0.5 Mb.

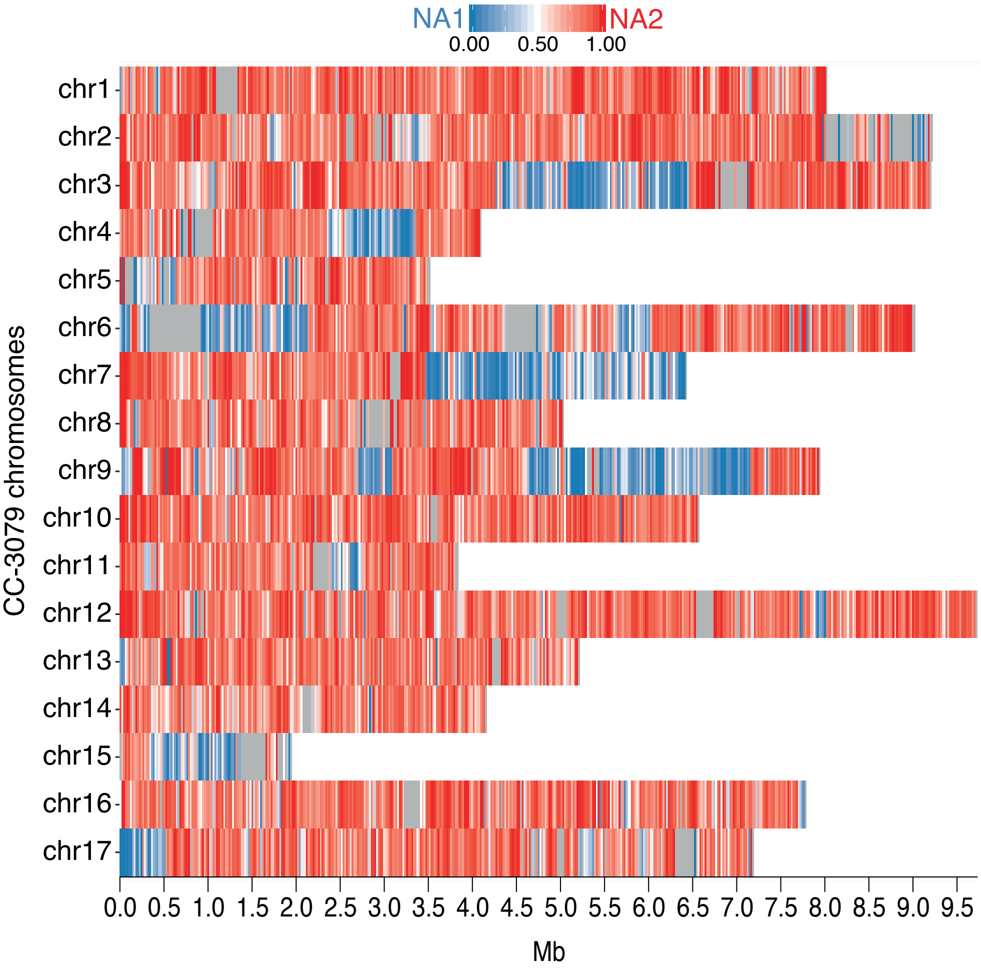


Figure S4. Admixture profiling for all chromosomes of the isolate CC-3079.

Appendix C

Supplementary Material for Chapter 3

The following supplementary notes and datasets are available from the Edinburgh Datashare repository with doi: https://doi.org/10.7488/ds/3103

Note S1. High molecular weight DNA extraction protocol for *Chlamydomonas.*

Dataset S1: Annotation notes for curated transposable elements from *Chlamydomonas incerta*, *Chlamydomonas schloesseri*, *Edaphochlamys debaryana*, *Eudorina* sp. and *Volvox carteri*. See Appendix E, Dataset S1 for *Chlamydomonas reinhardtii* transposable element annotation notes.

Dataset S2: Assembly metrics for genome assemblies of all available *Reinhardtinia* species and selected outgroups.

Dataset S3: Gene model annotation metrics for all available *Reinhardtinia* species and selected outgroups.

Dataset S4. List of contigs terminating in telomeric repeats.

Dataset S5. Orthogroup and InterPro domain annotation for *Chlamydomonas*-specific orthogroups with annotated domains.

Dataset S6: Orthogroup and InterPro domain annotation for *E. debaryana* gene family expansions (log2-transformed ratios >1) and contractions (ratios <-1) .

Dataset S7: Presence-absence of *C. reinhardtii* MT^{-} genes in *C. incerta*, *C. schloesseri* and *E. debaryana.*

Dataset S8: New genes identified with significant blastp homology to proteins from the *C. reinhardtii* v4.3 annotation.

Dataset S9: Core-*Reinhardtinia* ultraconserved elements (elements >=50 bp, 100% conservation within *Chlamydomonas* and >=95% conservation across eight core-*Reinhardtinia* species).

Table S1. Pacific Biosciences sequencing output. Metrics were calculated after removal of putative contaminant reads.

Species	*Chlamydomonas incerta*	*Chlamydomonas schloesseri*	*Edaphochlamys debaryana*
Platform	Sequel	Sequel	Sequel
Library	20 kb shear, 15-50 kb size selection	20 kb shear, 15-50 kb size selection	20 kb shear, 15-50 kb size selection
Yield (Gb)	6.31	6.18	5.70
Number of reads	820,556	861,783	728,942
Mean read length (bp)	7,692	7,171	7,822
Read length N50 (bp)	13,706	12,517	13,459

Table S2. Illumina genomic DNA sequencing datasets.

Species	**Platform**	**Library**	**Yield (Gb)**	**Number of reads**	**Read length (bp)**	**Insert size (bp)**
Chlamydomonas incerta	HiSeq 2000	high GC PCR conditions	3.34	33,413,300	2 x 100	180
Chlamydomonas incerta	HiSeq 2000	high GC PCR conditions	3.36	33,615,504	2 x 100	180
Chlamydomonas incerta	HiSeq 2000	high GC PCR conditions	2.09	21,752,290	2 x 100	5000
Chlamydomonas incerta	HiSeq 2000	high GC PCR conditions	2.31	24,116,474	2 x 100	5000
Chlamydomonas schloesseri	HiSeq 2500	high GC PCR conditions	3.03	24,250,400	2 x 125	300
Chlamydomonas schloesseri	HiSeq 2500	PCR free	3.09	24,775,262	2 x 125	300
Chlamydomonas schloesseri	HiSeq 2500	PCR free	2.25	18,071,364	2 x 125	500
Chlamydomonas schloesseri	HiSeq 2500	PCR free	0.87	7,005,992	2 x 125	500
Edaphochlamys debaryana	HiSeq 4000	PCR free	6.17	41,252,652	2 x 150	300

Table S3. Best megablast hits for ribosomal and plastid marker genes of the undescribed species *Chlamydomonas* sp. 3112.

Gene	Best megablast hit (described species)	Accession	Percent identity (%)	Best megablast hit (undescribed isolate)	Accession	Percent identity (%)
18S rRNA	*Pleodorina starrii* gene for 18S rRNA, partial sequence, strain: NIES-1362	LC086359.1	99.14	*Chlamydomonas sp. YACCYB320* 18S ribosomal RNA gene, partial sequence	MH683848.1	100.00
atpB	*Colemanosphaera charkowiensis* plastid, complete genome	MH511733.1	91.56	NA	NA	NA
psaA	*Chlamydomonas zebra* SAG 8.72 chloroplast psaA gene for photosystem I P700 chlorophyll a apoprotein A1, partial cds	LC380301.1	96.42	NA	NA	NA
psaB	*Chlamydomonas zebra* SAG 8.72 chloroplast psaB gene for photosystem I P700 chlorophyll a apoprotein A2, partial cds	LC380328.1	96.99	NA	NA	NA
psbC	*Chlamydomonas zebra* SAG 8.72 chloroplast psbC gene for photosystem II CP43 apoprotein, partial cds	LC380355.1	96.51	NA	NA	NA
rbcL	*Chlamydomonas zebra* SAG 8.72 chloroplast rbcL gene for ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit, partial cds	LC380380.1	97.52	NA	NA	NA

Table S4. *Chlamydomonas reinhardtii* putative centromeric coordinates and repeat content.

Chromosome	Start	End	Ns (%)	All TE/ satellites (%)	*L1-1_CR/ ZeppL-1_cRei* (%)	*Dualen* (%)
chromosome_1	1,125,300	1,305,700	79.05	20.94	10.83	10.11
chromosome_2	2,576,195	2,632,100	32.19	64.74	53.80	0.00
chromosome_3	6,779,288	6,875,500	30.50	59.63	19.87	5.83
chromosome_4	792,800	1,062,475	15.92	76.61	33.21	23.89
chromosome_5	2,326,715	2,439,300	0.22	74.03	39.09	15.96
chromosome_6	4,468,200	4,724,700	36.48	63.38	51.38	9.58
chromosome_7	3,067,650	3,209,850	26.85	64.18	39.88	16.50
chromosome_8	3,065,600	3,160,300	11.10	72.68	15.47	34.46
chromosome_9	5,319,800	5,471,860	0.07	94.91	68.71	16.12
chromosome_10	3,514,600	3,625,100	9.32	89.62	33.11	35.42
chromosome_11	Unknown	Unknown	NA	NA	NA	NA
chromosome_12	6,578,800	6,799,650	20.61	75.35	42.27	17.86
chromosome_13	4,225,970	4,315,800	49.87	50.02	28.87	11.02
chromosome_14	2,078,500	2,157,100	0.00	99.90	81.88	0.00
chromosome_15	Unknown	Unknown	NA	NA	NA	NA
chromosome_16	3,237,200	3,440,300	12.02	85.66	50.62	15.41
chromosome_17	6,340,900	6,600,350	54.97	40.24	25.84	3.61

Ns (%) is the percentage of unknown bases per centromere. *Dualen* (%) is total repeat content of all families of *Dualen* LINE elements, which constitute the second most abundant category of putative centromeric repeat.
L1-1_CR and *ZeppL-1_cRei* are synonyms for the same transposable element family (Appendix E, Dataset S1).

Table S5. Illumina RNA-seq datasets.

Species	*Chlamydomonas incerta*	*Chlamydomonas schloesseri*	*Edaphochlamys debaryana*
Platform	HiSeq X	HiSeq X	HiSeq X
Library	TruSeq stranded	TruSeq stranded	TruSeq stranded
Yield (Gb)	8.20	7.42	7.48
Number of reads	54,641,386	74,151,920	74,758,226
Read length (bp)	2 x 150	2 x 100	2 x 100
Insert size (bp)	325	256	279

Table S6. Programs and command line options for *Chlamydomonas incerta* genome assembly.

Program	Version	Command line options	Reference
miniasm	0.3-r179		Li (2016)
Blobtools	v1.0		Laetsch and Blaxter (2017)

Canu	1.7.1	genomeSize=130.8m
		correctedErrorRate=0.065 corMhapSensitivity=normal
			Koren et al. (2017)

pbalign	0.3.1		https://github.com/PacificBiosciences/pbalign
Arrow	2.3.2		https://github.com/PacificBiosciences/GenomicConsensus
bbduk.sh	38.16	ktrim=r k=23 mink=11 hdist=1 tbo	https://jgi.doe.gov/data-and-tools/bbtools/
bbmerge-auto.sh	38.16		https://jgi.doe.gov/data-and-tools/bbtools/
BWA-MEM	0.7.17-r1188		Li and Durbin (2009)
samtools	1.9		Li et al. (2009)
picard MarkDuplicates	2.18.11-SNAPSHOT	REMOVE_DUPLICATES=true	http://broadinstitute.github.io/picard/
Nxtrim	v0.4.3-6eb8d5e	--rf	O'Connell et al. (2015)
trimmomatic	0.38	PE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:3 MINLEN:25	Bolger et al. (2014)
STAR	STAR_2.6.1a	--twopassMode Basic	Dobin et al. (2013)
Pilon	1.22	--fix bases	Walker et al. (2014)
IGV	2.7.2		Robinson et al. (2011)
Circlator			Hunt et al. (2015)

Table S7. Programs and command line options for *Chlamydomonas schloesseri* genome assembly.

Program	Version	Command line options	Reference
miniasm	0.3-r179		Li (2016)
Blobtools	v1.0		Laetsch and Blaxter (2017)

Canu	1.7.1	genomeSize=130.5m
		correctedErrorRate=0.065 corMhapSensitivity=normal
			Koren et al. (2017)

pbalign	0.3.1		https://github.com/PacificBiosciences/pbalign
Arrow	2.3.2		https://github.com/PacificBiosciences/GenomicConsensus
bbduk.sh	38.16	ktrim=r k=23 mink=11 hdist=1 tbo	https://jgi.doe.gov/data-and-tools/bbtools/
BWA-MEM	0.7.17-r1188		Li and Durbin (2009)
samtools	1.9		Li et al. (2009)
picard MarkDuplicates	2.18.11-SNAPSHOT	REMOVE_DUPLICATES=true	http://broadinstitute.github.io/picard/
trimmomatic	0.38	PE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:3 MINLEN:25	Bolger et al. (2014)
STAR	STAR_2.6.1a	--twopassMode Basic	Dobin et al. (2013)
Pilon	1.22	--fix bases	Walker et al. (2014)
IGV	2.7.2		Robinson et al. (2011)
Circlator			Hunt et al. (2015)

Table S8. Programs and command line options for *Edaphochlamys debaryana* genome assembly.

Program	Version	Command line options	Reference
miniasm	0.3-r179		Li (2016)
Blobtools	v1.0		Laetsch and Blaxter (2017)

Canu	1.7.1	genomeSize=148.9m
		correctedErrorRate=0.065 corMhapSensitivity=normal
			Koren et al. (2017)

pbalign	0.3.1		https://github.com/PacificBiosciences/pbalign
Arrow	2.3.2		https://github.com/PacificBiosciences/GenomicConsensus
bbduk.sh	38.16	ktrim=r k=23 mink=11 hdist=1 tbo	https://jgi.doe.gov/data-and-tools/bbtools/
BWA-MEM	0.7.17-r1188		Li and Durbin (2009)
samtools	1.9		Li et al. (2009)
picard MarkDuplicates	2.18.11-SNAPSHOT	REMOVE_DUPLICATES=true	http://broadinstitute.github.io/picard/
trimmomatic	0.38	PE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:3 MINLEN:25	Bolger et al. (2014)
STAR	STAR_2.6.1a	--twopassMode Basic	Dobin et al. (2013)
Pilon	1.22	--fix bases	Walker et al. (2014)
Circlator			Hunt et al. (2015)

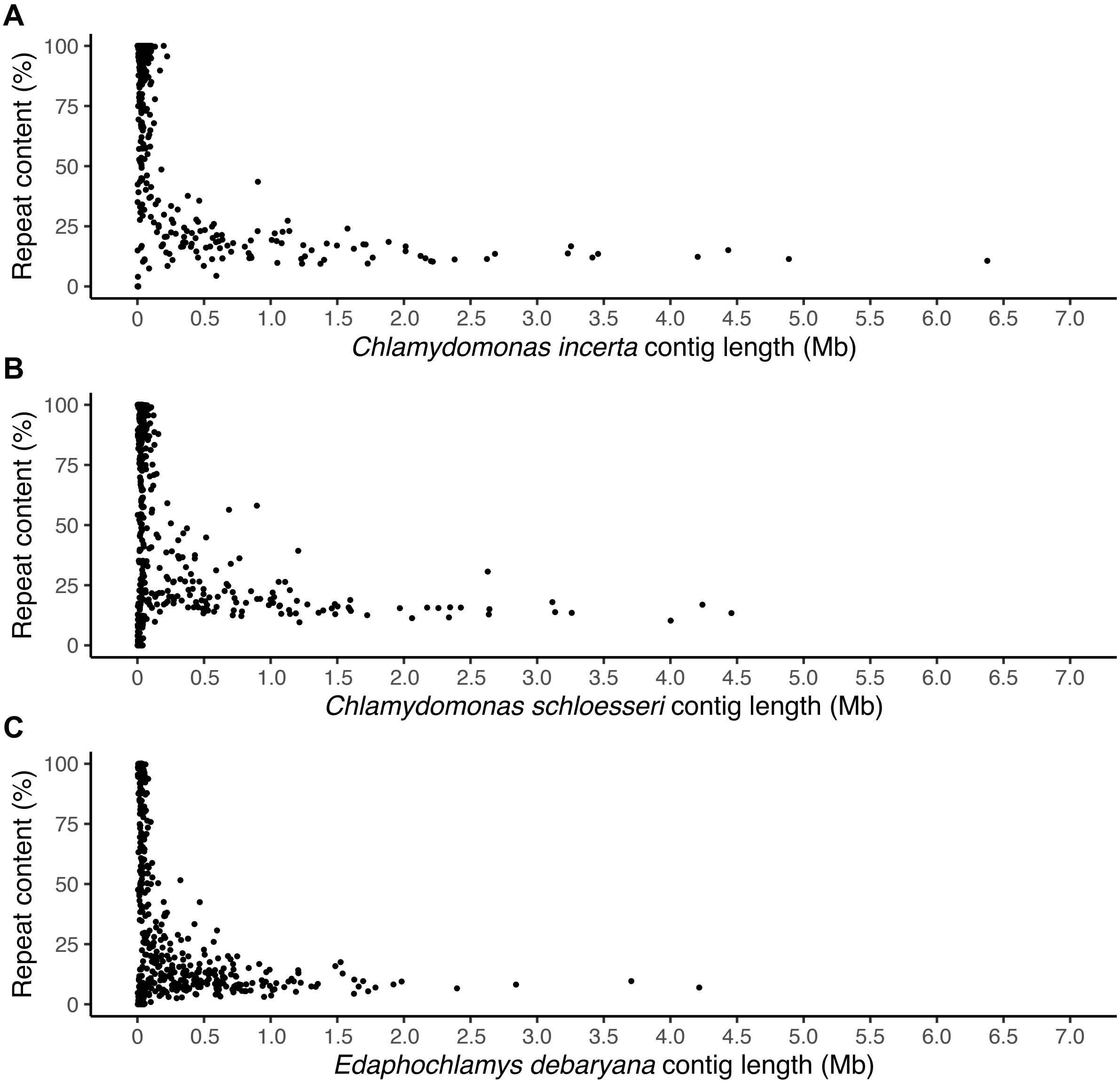


Figure S1. Total repeat content per contig (transposable elements, satellites and simple/low-complexity repeats) plotted by contig length.
(A) *Chlamydomonas incerta.*
(B) *Chlamydomonas schloesseri.*
(C) *Edaphochlamys debaryana.*

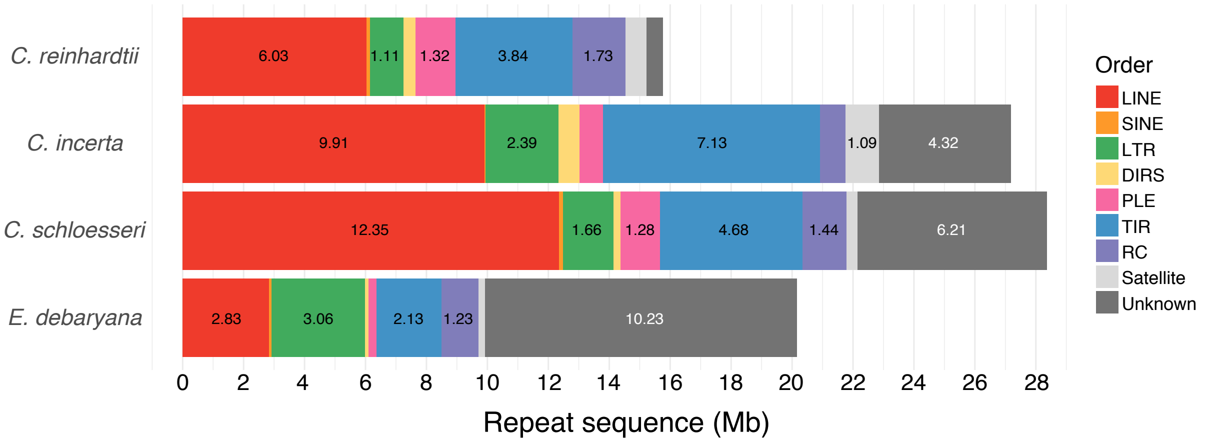


Figure S2. Repeat content per species by repeat order.
Numbers within bars represent total sequence per order in megabases. LINE = long interspersed nuclear element, SINE = short interspersed nuclear element, LTR = long terminal repeat, DIRS = tyrosine recombinase encoding retrotransposons, PLE = *Penelope*-like elements, TIR = terminal inverted repeat (i.e. DNA transposons), RC = rolling-circle elements. Note that the cumulative repeat content totals are marginally higher than those in Table 1 due to redundancy in repeat classification.

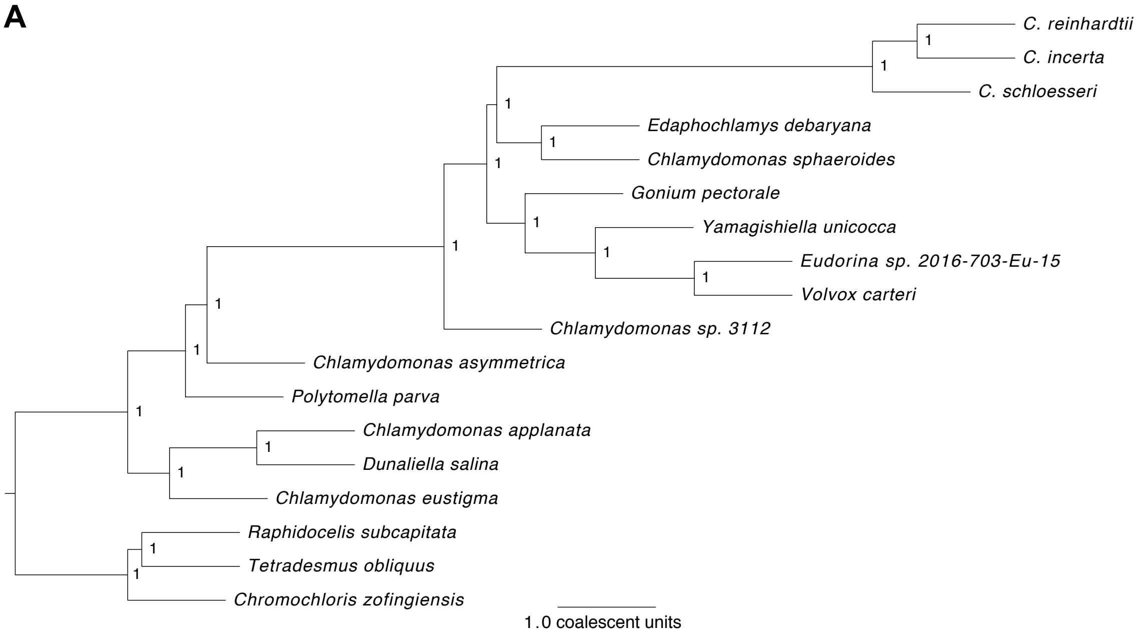


Figure S3. Phylogenomic analyses (continues next page).

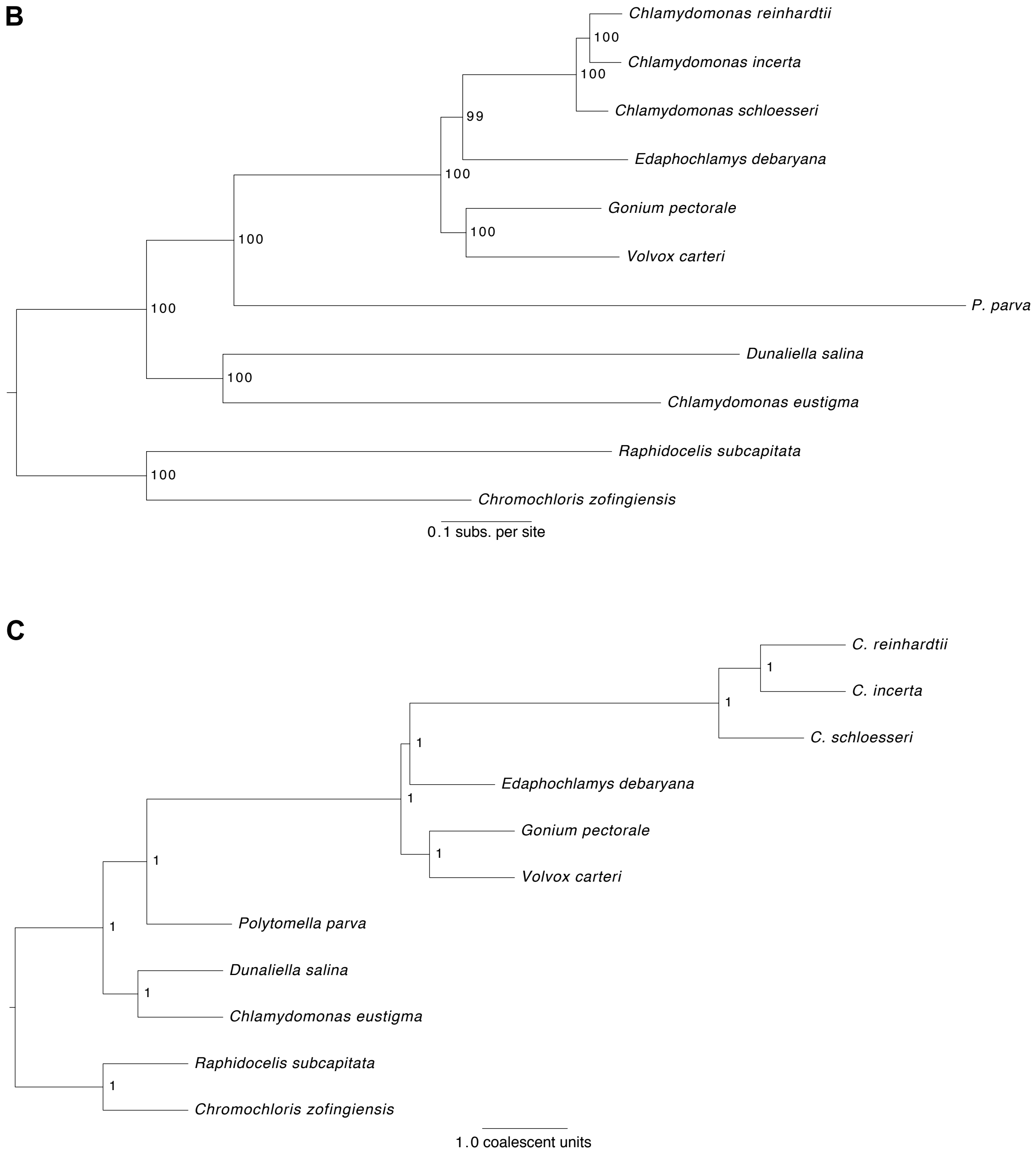


Figure S3. Phylogenomic analyses.
(A) ASTRAL-III species tree (15 Volvocales species and three outgroups) summarising 1,624 gene trees produced from individual protein alignments of chlorophyte BUSCO genes.
(B) ML phylogeny of nine Volvocales species and two outgroups inferred using LG+F+R5 model and a concatenated protein alignment of 1,681 putative single-copy orthologs identified by OrthoFinder.
(C) ASTRAL-III species tree summarising 1,681 gene trees produced from individual protein alignments of the OrthoFinder single-copy genes.
Note that support values for **(A)** and **(C)** represent local posterior probabilities, while **(B)** represents ultrafast bootstrap values.

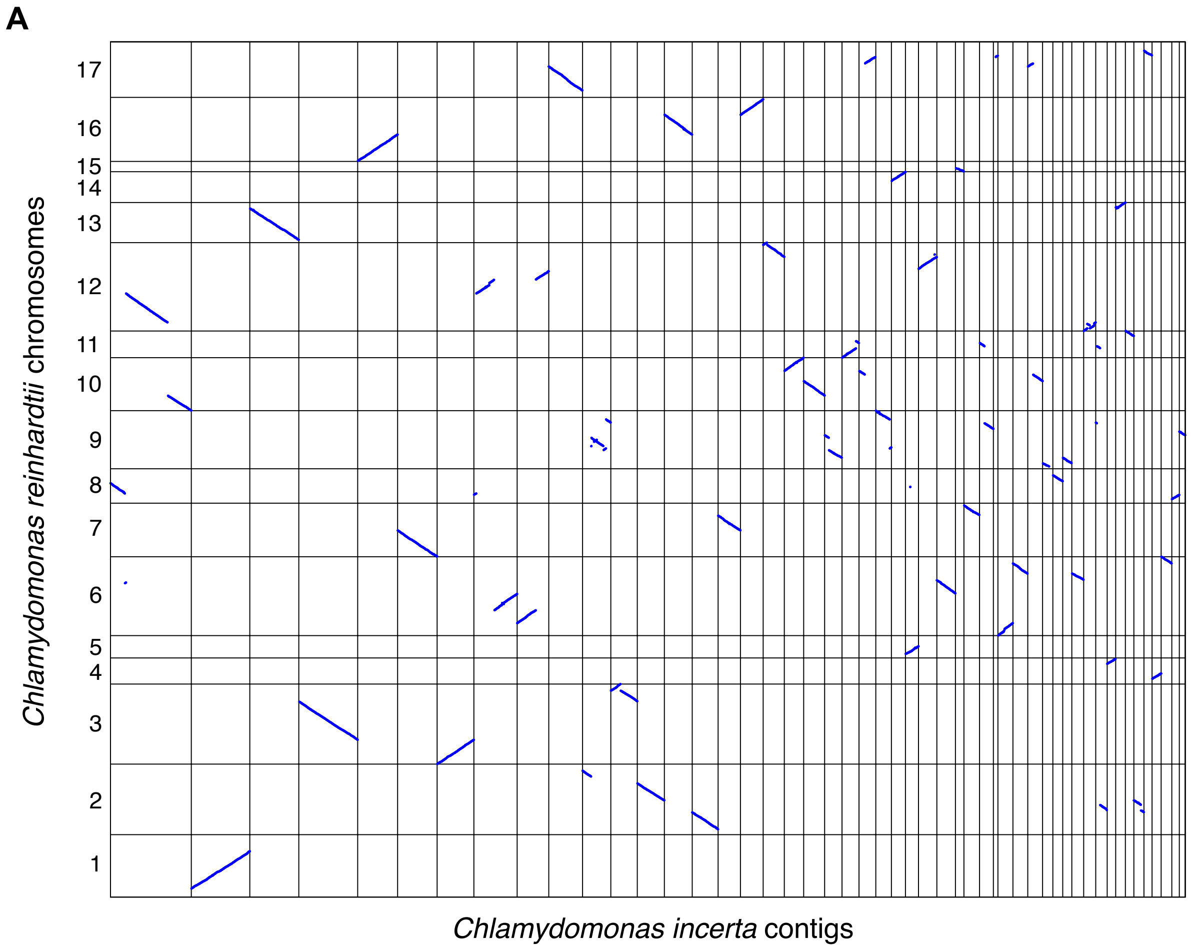


Figure S4. Dotplots representing syntenic genomic segments (continues next page).

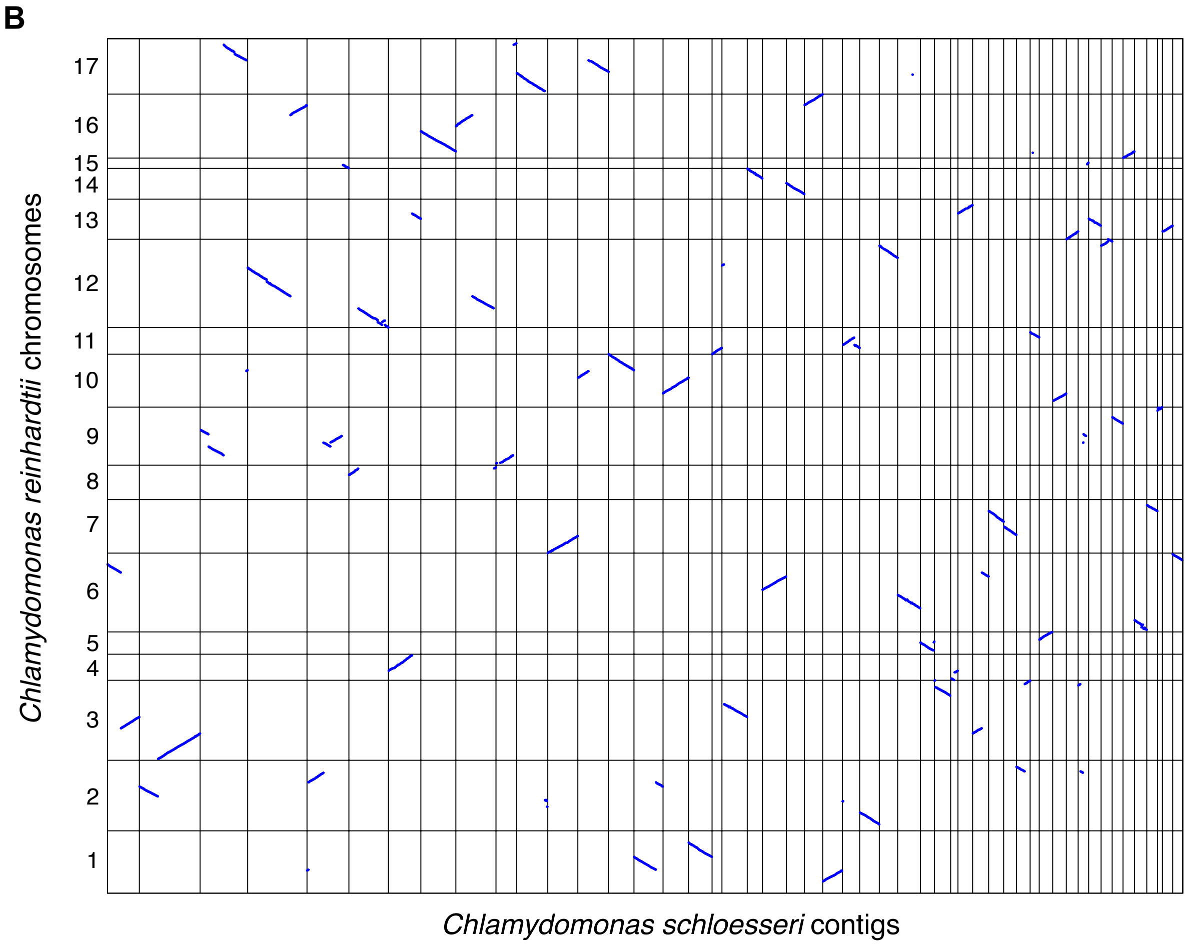


Figure S4. Dotplots representing syntenic genomic segments (continues next page).

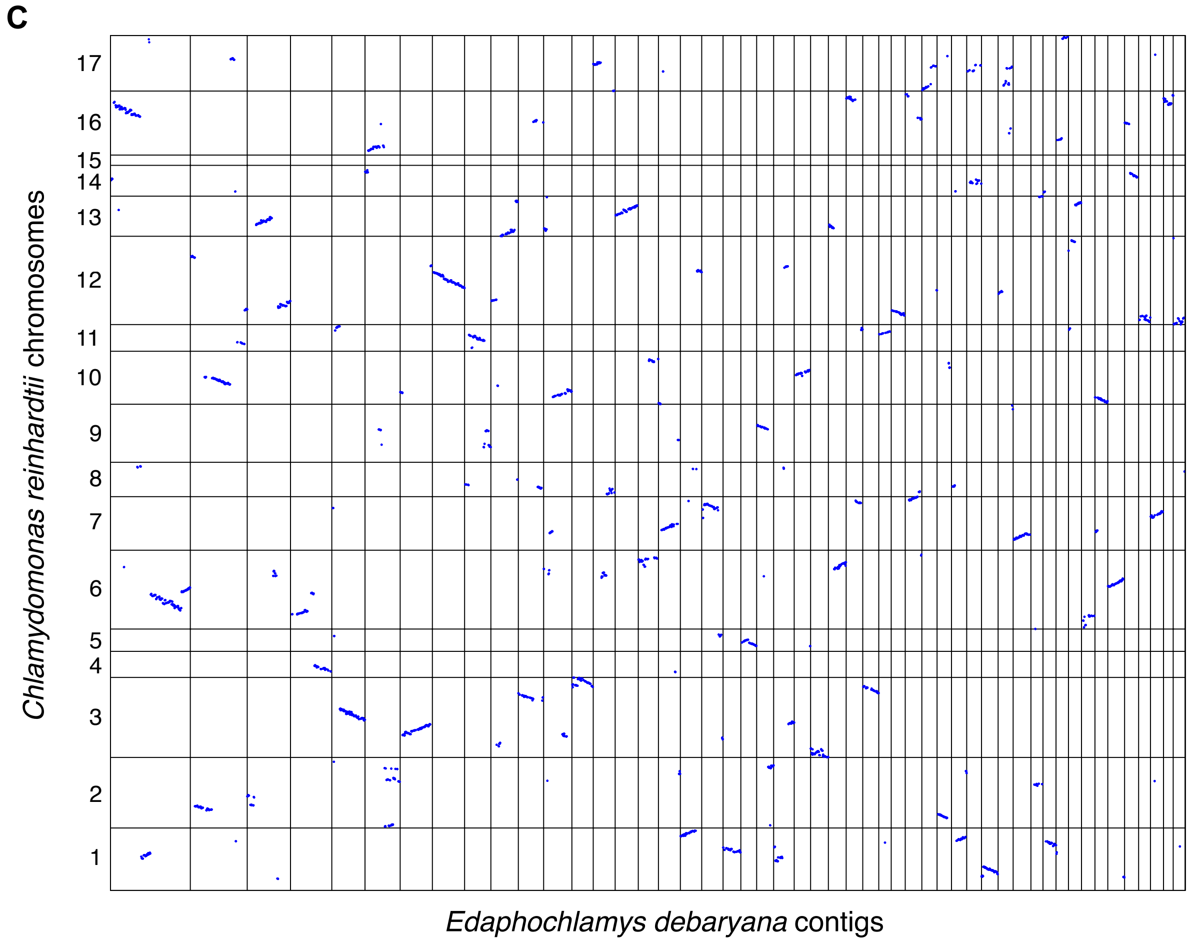


Figure S4. Dotplots representing syntenic genomic segments.
Each plot presents syntenic segments between *C. reinhardtii* and the 50 largest contigs of:
(A) *Chlamydomonas incerta.*
(B) *Chlamydomonas schloesseri.*
(C) *Edaphochlamys debaryana.*

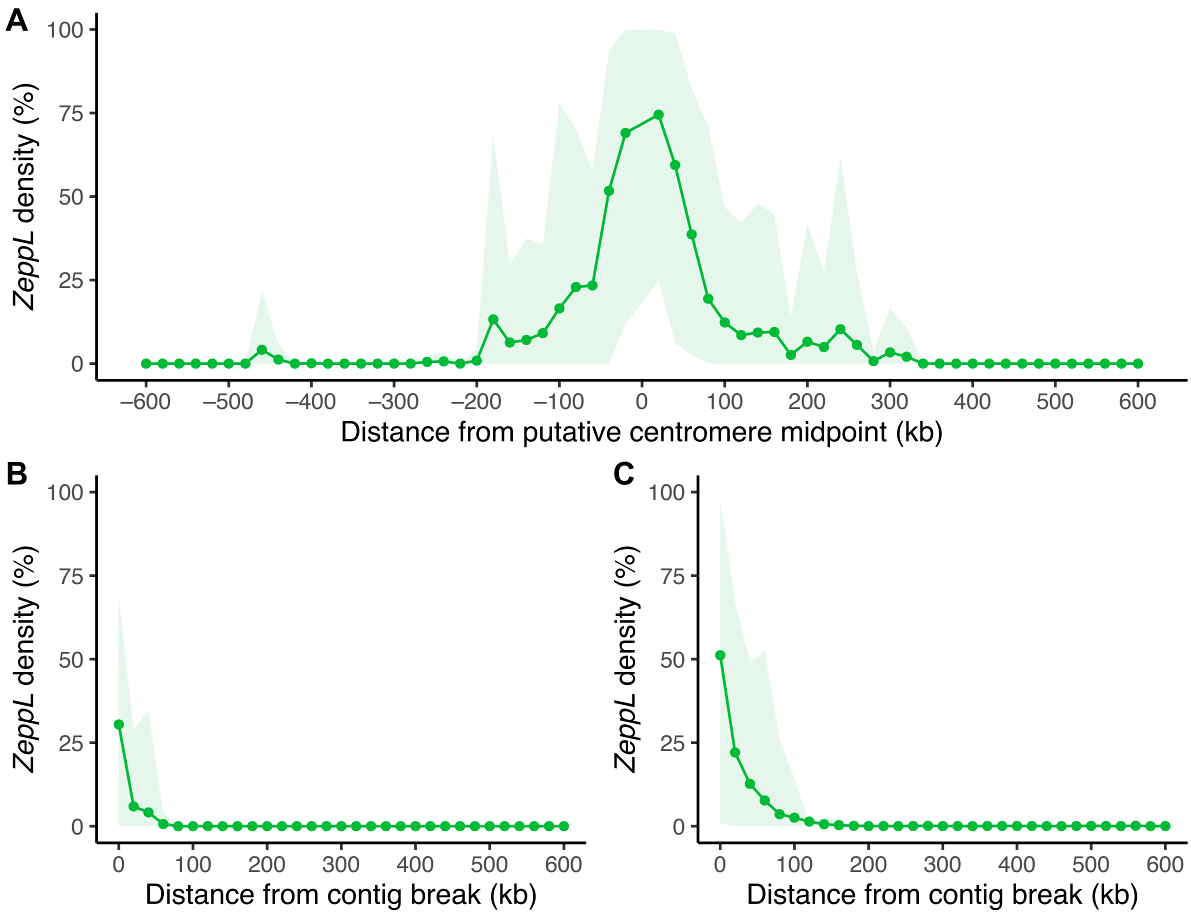


Figure S5. Mean densities of *Zepp*-like *L1* LINE elements per 20 kb windows averaged over relevant chromosomes/contigs. Shaded areas represent 95% quantiles.

(A) Density of *L1-1_CR* / *ZeppL-1_cRei* elements relative to midpoint of 15 putative *C. reinhardtii* centromeres.

(B) Density of *ZeppL-1_cInc* elements relative to *C. incerta* contig ends syntenic to *C. reinhardtii* putative centromeres.

(C) Density of *ZeppL-1_cSch* and *ZeppL-2_cSch* elements relative to *C. schloesseri* contig ends syntenic to *C. reinhardtii* putative centromeres.

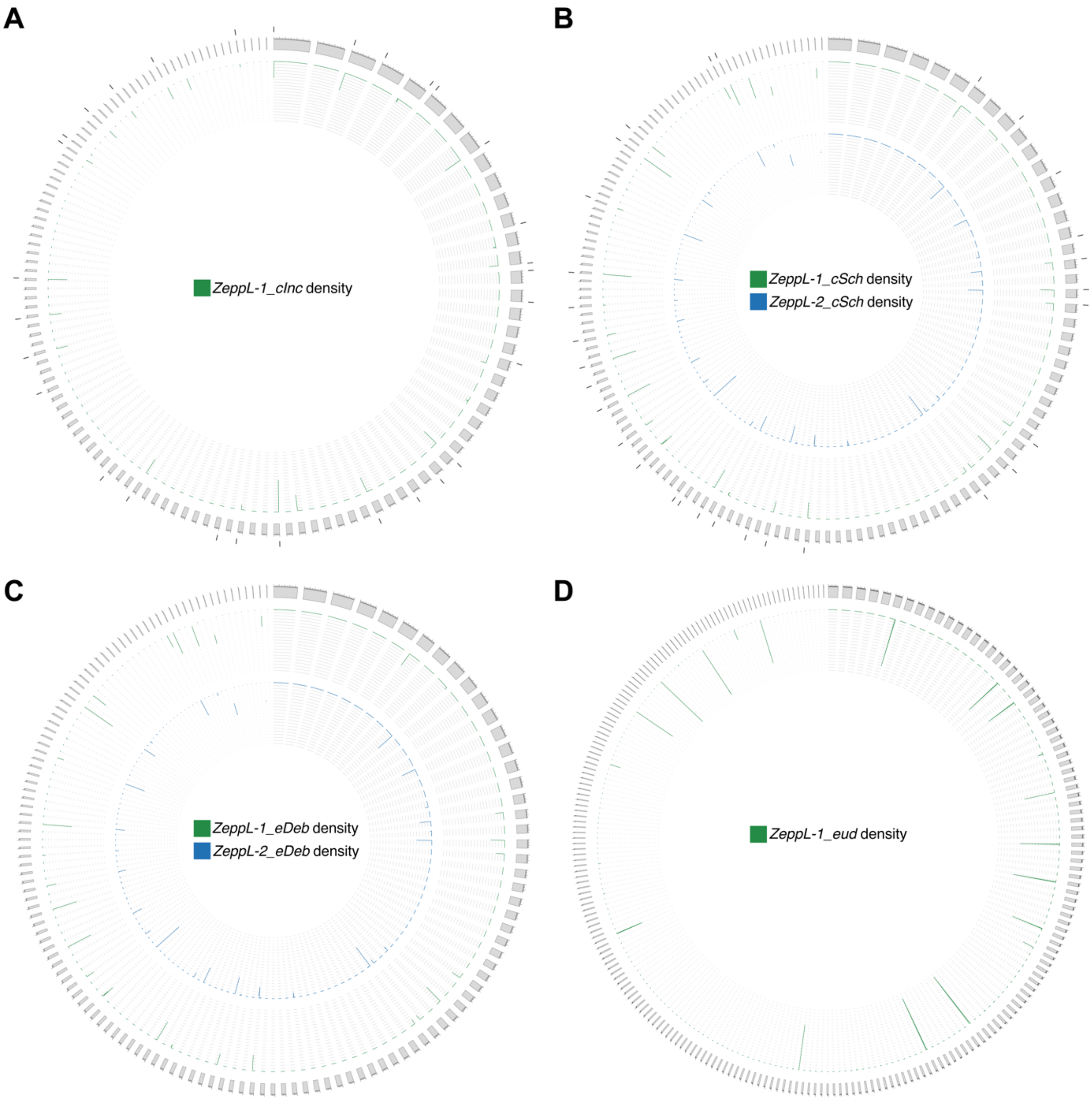


Figure S6. Genome-wide density of *Zepp*-like elements.
Contigs are represented by grey bands and ordered by size. Dark grey ticks above/below contigs represent contig ends inferred as syntenic with *C. reinhardtii* centromeres. Axis ranges from 0-100% and densities calculated for 50 kb windows.
(A) *Chlamydomonas incerta*.
(B) *Chlamydomonas schloesseri*.
(C) *Edaphochlamys debaryana*.
(D) *Eudorina* sp. 2016-703-Eu-15.

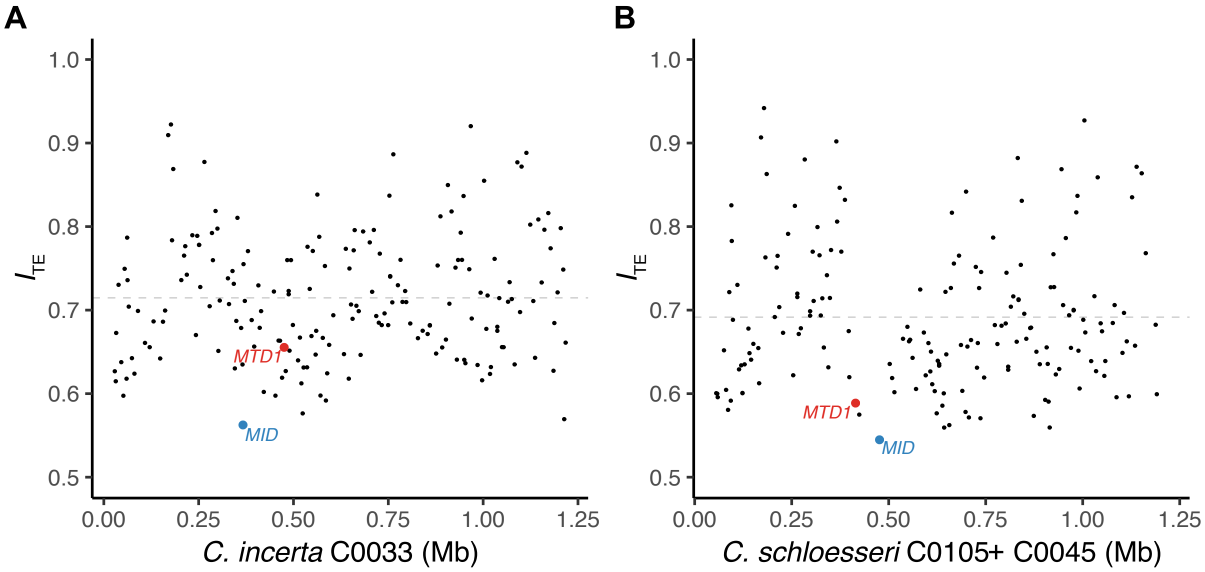


Figure S7. Codon adaptation of *minus* mating type genes.

I_{TE} values are plotted for each gene across the contigs containing the putative minus mating type loci of *C. incerta* **(A)** and *C. schloesseri* **(B).** Each point represents a gene, with *MID* and *MTD1* orthologs highlighted. Dashed grey lines represent genome-wide means. Note that for *C. schloesseri* the region syntenous to the *C. reinhardtii* mating type is entirely on contig C0045, C0105 was appended to C0045 to show the genes syntenous with the most telomere-proximal region of *C. reinhardtii* chromosome 6.

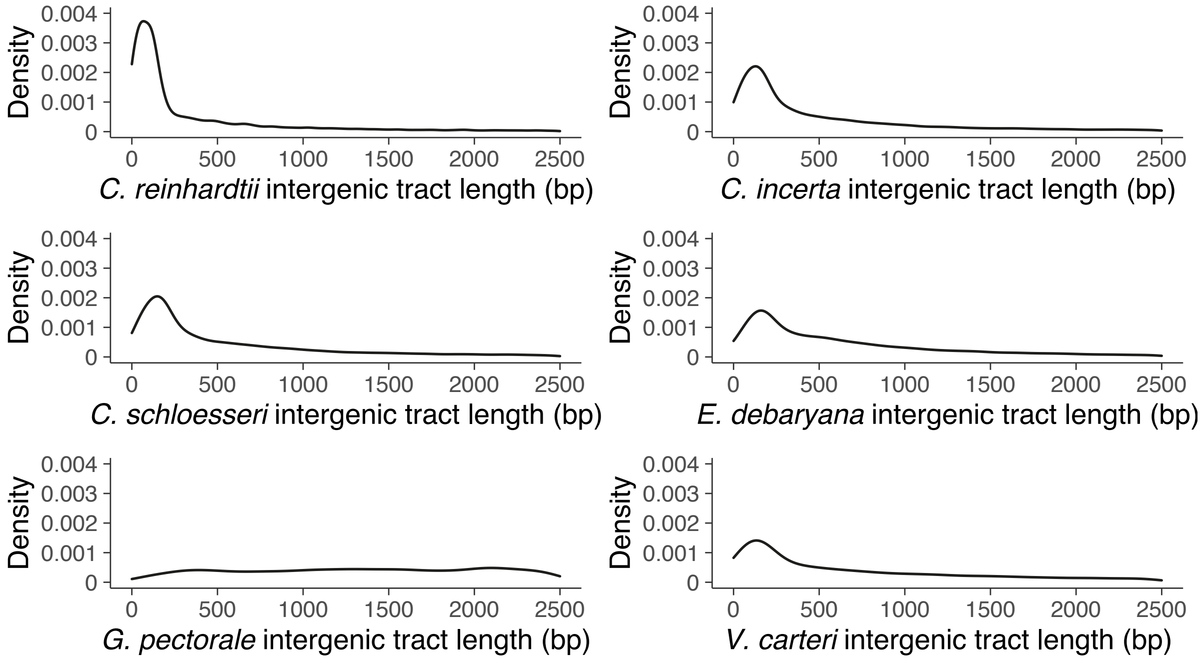


Figure S8. Distribution of intergenic tract lengths across six core-*Reinhardtinia* species. The *G. pectorale* distribution likely differs due to the lack of UTR annotation for this species.

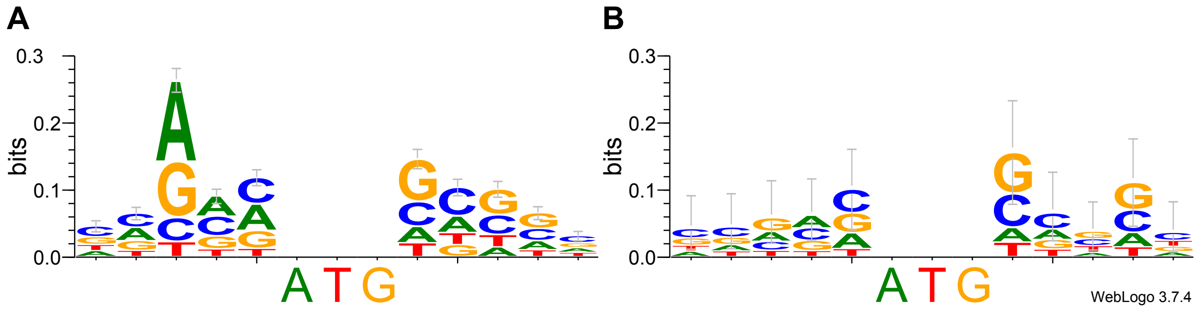


Figure S9. Kozak consensus sequence logos.
(A) A randomly selected half of the control gene set.
(B) The 250 low coding potential genes that failed all three coding potential analyses.

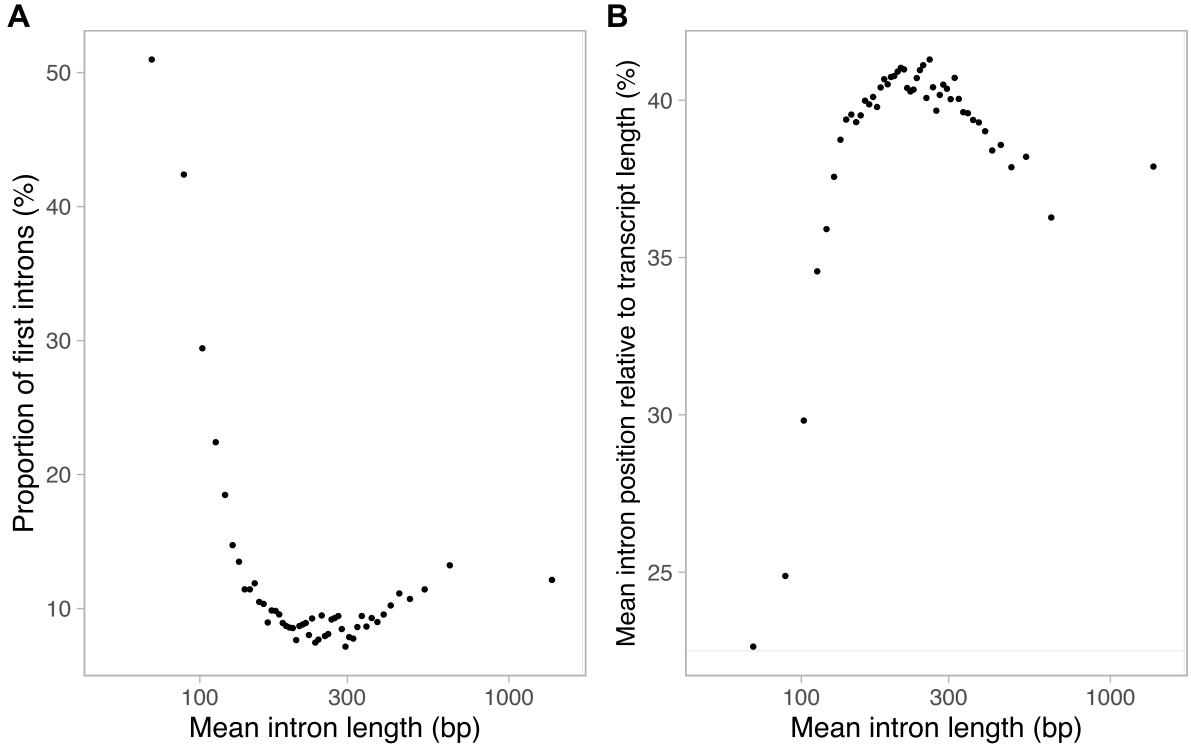


Figure S10. Relationship between intron lengths and intron locations within genes.
(A) Relationship between the proportion of introns that are the first intron of a gene and the mean intron length per bin (3.4.9).
(B) The relationship between the mean intron position relative to transcript length (e.g. an intron at position 500 of a 2000 bp transcript equals 25%) and mean intron length per bin.

Appendix D

Supplementary Material for Chapter 4

The following supplementary datasets are available from the Edinburgh Datashare repository with doi: https://doi.org/10.7488/ds/3103

Dataset S1: Metrics and sequence context of all CC-503 v6 assembly gaps.

Dataset S2: Curated structural mutations unique to the CC-503 v6 assembly.

Dataset S3: Curated structural mutations unique to the CC-4532 v6 assembly.

Dataset S4. Metrics and sequence context of all CC-4532 v6 assembly gaps.

Dataset S5. Curated TE transposition events unique to the CC-503 v6 assembly.

Dataset S6: Curated TE transposition events unique to the CC-4452 v6 assembly.

Table S1. Summary statistics, gene density and repeat content of the CC-503 v6 chromosomes.

Chromosome	Length (bp)	Ns (%)	Gene density (%)	All repeats (%)	TEs (%)	Microsatellite (%)	Satellite (%)
chromosome_01	7,951,217	1.46	82.50	13.72	8.52	1.74	3.45
chromosome_02	9,475,574	0.13	82.07	11.55	7.19	1.45	2.90
chromosome_03	9,058,378	0.71	83.37	12.40	8.08	1.23	3.08
chromosome_04	4,047,166	0.24	72.82	21.01	13.96	2.43	4.62
chromosome_05	3,645,384	1.74	75.17	21.84	13.68	2.87	5.28
chromosome_06	8,851,674	1.14	80.18	12.88	8.42	1.34	3.12
chromosome_07	6,360,637	0.15	81.77	13.89	9.15	1.76	2.98
chromosome_08	4,474,995	0.94	80.28	15.59	8.88	2.18	4.54
chromosome_09	5,727,556	0.55	78.26	15.93	10.69	1.85	3.39
chromosome_10	6,696,547	1.13	82.87	13.19	8.97	1.44	2.78
chromosome_11	4,467,042	0.85	78.26	17.54	10.61	1.94	4.99
chromosome_12	9,698,580	1.75	81.32	12.61	8.73	1.41	2.46
chromosome_13	5,201,819	1.30	82.87	13.50	7.71	1.99	3.80
chromosome_14	4,101,833	1.95	82.30	15.46	8.41	2.21	4.84
chromosome_15	5,632,506	15.37	43.38	45.27	35.36	2.84	7.07
chromosome_16	7,825,789	0.20	80.64	14.78	9.93	1.40	3.45
chromosome_17	6,881,933	1.36	77.84	15.68	10.87	1.53	3.28
unplaced contigs	1,448,066	0.00	9.36	77.52	45.92	3.64	27.97

Table S2. CC-503 v6 genomic site classes and overlap by repetitive sequences.

Site class	CDS	5' UTR	3' UTR	intronic	intergenic <250 bp	intergenic ≥250 bp
Total sequence (Mb)	37.48	3.88	9.90	34.38	0.39	23.67
Genomic proportion (%)	34.17	3.53	9.03	31.34	0.35	21.58
TEs (%)	0.40	3.34	7.00	4.80	2.52	39.34
Proportion total TEs (%)	1.24	1.08	5.80	13.83	0.08	77.96
Microsatellite (%)	0.92	0.31	0.40	3.46	0.22	1.44
Proportion total microsatellite (%)	17.83	0.63	2.06	61.77	0.04	17.66
Satellite (%)	3.21	0.91	1.22	4.68	0.61	5.70
Proportion total satellite (%)	27.86	0.82	2.79	37.24	0.05	31.23

Table S3. Putative centromere metrics of the CC-1690 and CC-503 v6 assemblies.

	CC-1690				**CC-503 v6**					
Chromosome	**Cluster start (bp)**	**Cluster end (bp)**	**Length (bp)**	***ZeppL-1_cRei*** **(%)**	**Cluster start (bp)**	**Cluster end (bp)**	**Length (bp)**	**Ns (%)**	***ZeppL-1_cRei*** **(%)**	**Proportion assembled (%)**
chromosome_01	1,011,599	1,292,049	280,450	61.09	1,030,221	1,301,561	271,340	33.30	49.46	64.53
chromosome_02	2,537,384	2,612,189	74,805	73.96	2,544,784	2,617,649	72,865	16.86	73.49	80.99
chromosome_03	6,818,820	7,042,755	223,935	68.08	6,790,993	7,014,973	223,980	26.84	58.18	73.18
chromosome_04	818,524	1,046,741	228,217	52.37	820,430	1,055,411	234,981	3.89	51.30	98.95
chromosome_05	1,585,545	1,723,249	137,704	67.19	1,594,970	1,732,727	137,757	44.15	57.68	55.87
chromosome_06	4,421,663	4,666,803	245,140	59.78	4,386,292	4,632,001	245,709	14.03	53.30	86.17
chromosome_07	3,001,398	3,162,084	160,686	46.32	2,967,574	3,127,402	159,828	6.07	43.58	93.43
chromosome_08	2,633,327	2,744,878	111,551	77.34	2,605,428	2,718,394	112,966	37.08	66.13	63.72
chromosome_09	4,174,093	4,350,962	176,869	62.31	3,158,199	3,334,901	176,702	17.86	54.41	82.06
chromosome_10	3,577,834	3,775,024	197,190	65.75	3,552,180	3,740,237	188,057	38.49	50.83	58.66
chromosome_11	2,899,970	3,069,384	169,414	59.55	2,823,909	2,970,579	146,670	0.00	31.90	86.57
chromosome_12	6,619,662	6,885,102	265,440	59.80	6,584,224	6,838,818	254,594	26.55	43.96	70.45
chromosome_13	4,296,370	4,384,338	87,968	69.53	4,224,607	4,256,944	32,337	0.00	24.47	36.76
chromosome_14	2,108,172	2,226,090	117,918	86.18	2,099,549	2,217,501	117,952	46.75	74.37	53.27
chromosome_15	3,334,962	3,773,685	438,723	34.27	3,231,633	3,670,670	439,037	27.51	15.91	72.54
chromosome_16	3,246,217	3,499,602	253,385	63.30	3,199,601	3,445,113	245,512	3.44	61.95	93.56
chromosome_17	5,970,579	6,293,255	322,676	54.59	5,966,525	6,289,324	322,799	26.16	38.65	73.87

Proportion assembled is estimated relative to the length of *ZeppL* clusters in CC-1690. Note that although chromosome 11 and 13 clusters are shown as 0% Ns there are gaps immediately following the final *ZeppL-1_cRei* elements that presumably contain centromeric sequence that is assembled in CC-1690.

Table S4. Summary statistics, gene density and repeat content of the CC-4532 v6 chromosomes.

Chromosome	Length (bp)	Ns (%)	Gene density (%)	All repeats (%)	TEs (%)	Microsatellite (%)	Satellite (%)
chromosome_01	8,225,636	0.39	80.24	15.88	10.28	1.71	3.89
chromosome_02	8,655,884	0.23	83.59	11.95	7.98	1.35	2.62
chromosome_03	9,286,894	0.24	82.24	14.26	9.90	1.26	3.10
chromosome_04	4,130,073	0.53	73.05	22.11	14.88	2.43	4.79
chromosome_05	3,682,160	1.03	74.90	23.56	15.67	2.75	5.14
chromosome_06	8,913,359	0.30	79.63	14.87	10.25	1.32	3.30
chromosome_07	6,492,107	0.17	80.78	15.11	10.31	1.88	2.92
chromosome_08	4,526,983	0.31	79.48	16.50	9.94	2.19	4.37
chromosome_09	6,807,148	0.59	75.41	17.51	12.06	1.86	3.58
chromosome_10	6,800,247	0.13	81.04	15.56	11.24	1.55	2.76
chromosome_11	4,479,522	0.29	78.05	18.66	11.83	1.92	4.91
chromosome_12	9,952,739	0.66	79.68	14.41	10.56	1.36	2.49
chromosome_13	5,281,438	0.11	81.78	15.46	8.91	1.92	4.63
chromosome_14	4,217,303	0.00	80.95	18.02	10.81	2.12	5.10
chromosome_15	5,870,643	9.19	40.44	43.62	33.70	2.85	7.07
chromosome_16	8,042,475	0.24	79.53	15.99	10.85	1.42	3.73
chromosome_17	6,954,842	0.62	76.67	17.05	12.52	1.47	3.05
unplaced contigs	1,716,047	0.00	18.88	66.74	33.15	4.05	29.54

Table S5. Putative centromere metrics of the CC-1690 and CC-4532 v6 assemblies.

	CC-1690				CC-4532 v6					
Chromosome	**Cluster start (bp)**	**Cluster end (bp)**	**Length (bp)**	***ZeppL-1_cRei* (%)**	**Cluster start (bp)**	**Cluster end (bp)**	**Length (bp)**	**Ns (%)**	***ZeppL-1_cRei* (%)**	**Proportion assembled (%)**
chromosome_01	1,011,599	1,292,049	280,450	61.09	1,041,556	1,302,925	261,369	11.43	59.61	82.55
chromosome_02	2,537,384	2,612,189	74,805	73.96	2,591,685	2,676,089	84,404	9.08	75.14	102.59
chromosome_03	6,818,820	7,042,755	223,935	68.08	6,947,172	7,186,274	239,102	5.04	55.58	101.39
chromosome_04	818,524	1,046,741	228,217	52.37	834,190	1,059,835	225,645	0.00	51.88	98.87
chromosome_05	1,585,545	1,723,249	137,704	67.19	1,594,574	1,735,343	140,769	0.00	68.55	102.23
chromosome_06	4,421,663	4,666,803	245,140	59.78	4,356,487	4,621,046	264,559	0.04	57.02	107.88
chromosome_07	3,001,398	3,162,084	160,686	46.32	3,022,198	3,198,066	175,868	0.00	42.33	109.45
chromosome_08	2,633,327	2,744,878	111,551	77.34	2,633,797	2,747,480	113,683	12.20	75.11	89.48
chromosome_09	4,174,093	4,350,962	176,869	62.31	4,193,959	4,371,195	177,236	0.00	62.36	100.21
chromosome_10	3,577,834	3,775,024	197,190	65.75	3,582,600	3,786,653	204,053	4.15	65.46	99.18
chromosome_11	2,899,970	3,069,384	169,414	59.55	2,864,587	3,025,130	160,543	0.69	57.00	94.11
chromosome_12	6,619,662	6,885,102	265,440	59.80	6,642,004	6,922,583	280,579	1.84	55.63	103.76
chromosome_13	4,296,370	4,384,338	87,968	69.53	4,317,677	4,416,966	99,289	5.60	21.70	106.54
chromosome_14	2,108,172	2,226,090	117,918	86.18	2,181,897	2,290,737	108,840	0.00	85.04	92.30
chromosome_15	3,334,962	3,773,685	438,723	34.27	3,414,857	3,866,597	451,740	3.71	30.77	99.15
chromosome_16	3,246,217	3,499,602	253,385	63.30	3,326,275	3,581,730	255,455	0.00	63.46	100.82
chromosome_17	5,970,579	6,293,255	322,676	54.59	6,025,440	6,345,454	320,014	11.35	48.34	87.92

Proportion assembled is relative to length of *ZeppL* clusters in CC-1690 assembly. Note that clusters in CC-4532 v6 are often longer than in CC-1690 as a result of novel TE insertions of *Gypsy-7a_cRei*.

Table S6. Genomic coordinates of "Haplotype 2" in the CC-4532 v6 assembly.

CC-4532 chromosome	CC-4532 start	CC-4532 end	Length (kb)
chromosome_03	8,636,170	8,789,722	153.55
chromosome_06	1	976,104	976.10
chromosome_06	1,153,483	1,862,040	708.56
chromosome_11	1,342,177	2,225,240	883.06
chromosome_15	979,323	1,425,187	445.86
chromosome_17	3,921,106	5,787,376	1866.27

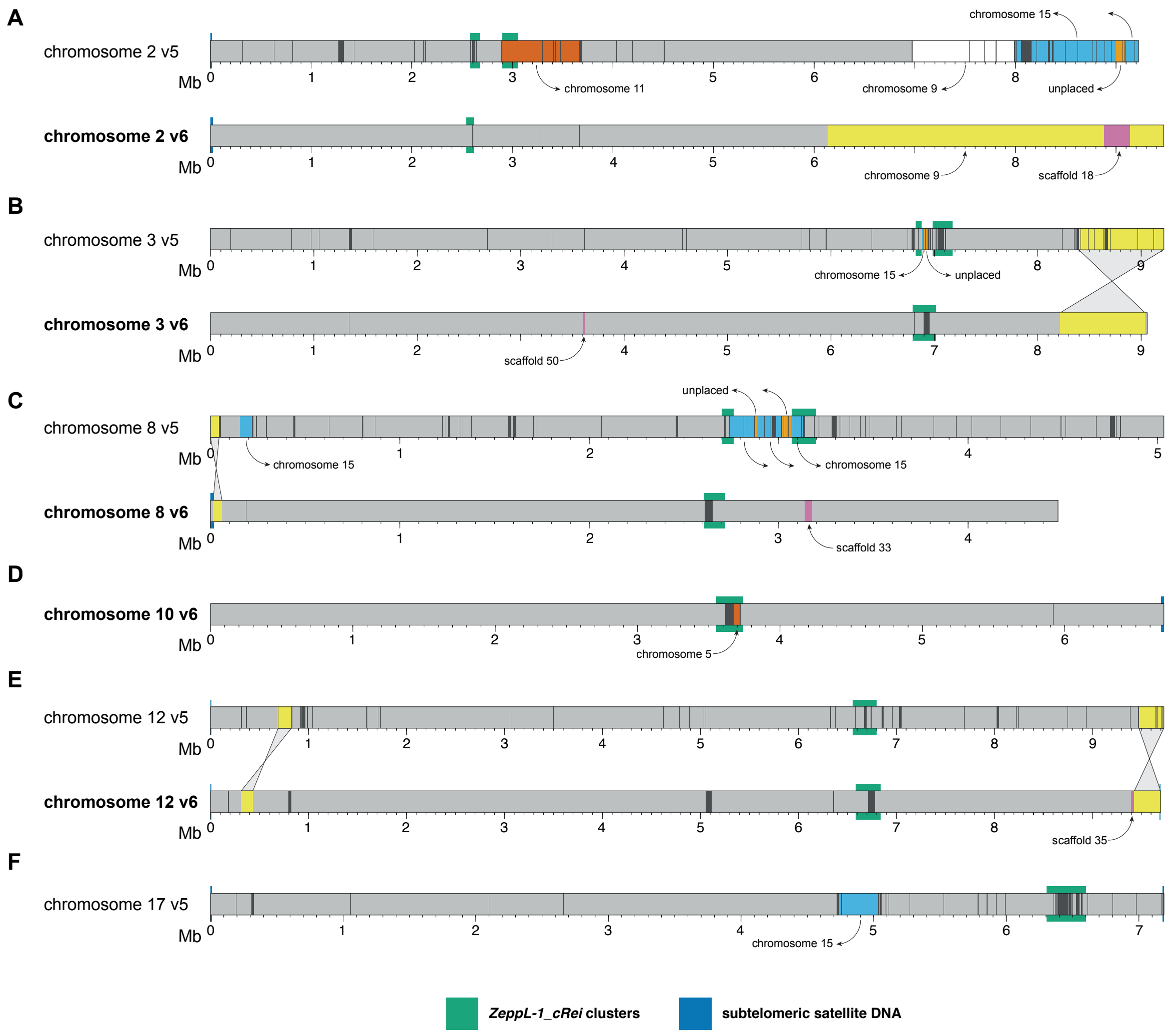


Figure S1. Misassemblies in v5 and their resolution in CC-503 v6.
Overview of all chromosomal changes between the assemblies. Chromosomes 5 and 11 are shown in Figure 2. Chromosome 9 is not shown since the only change relates to the reciprocal translocation shown in Figure 4.
(A) Chromosome 2 (see Figure 4 for translocation information).
(B) Chromosome 3.
(C) Chromosome 8.
(D) Chromosome 10 (v5 not shown since only change is the movement of a single sequence from chromosome 5).
(E) Chromosome 12.
(F) Chromosome 17 (CC-503 v6 not shown since only change is the movement of a single sequence to chromosome 15).

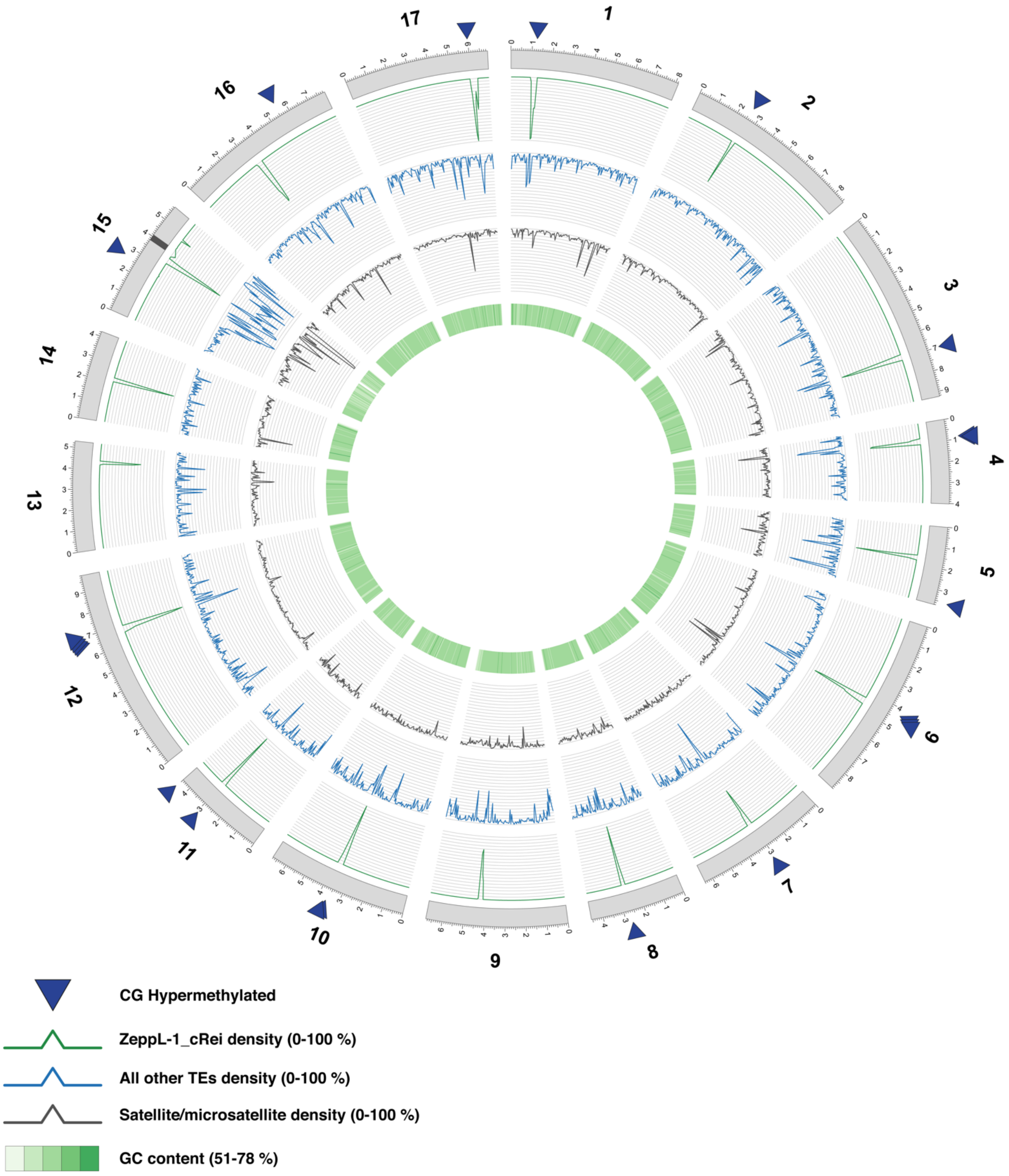


Figure S2. The CC-1690 genome assembly.
Circos plot representation of CC-1690. Grey outer blocks represent chromosomes and dark grey regions represent gaps between contigs. All metrics were calculated for 50 kb non-overlapping windows.

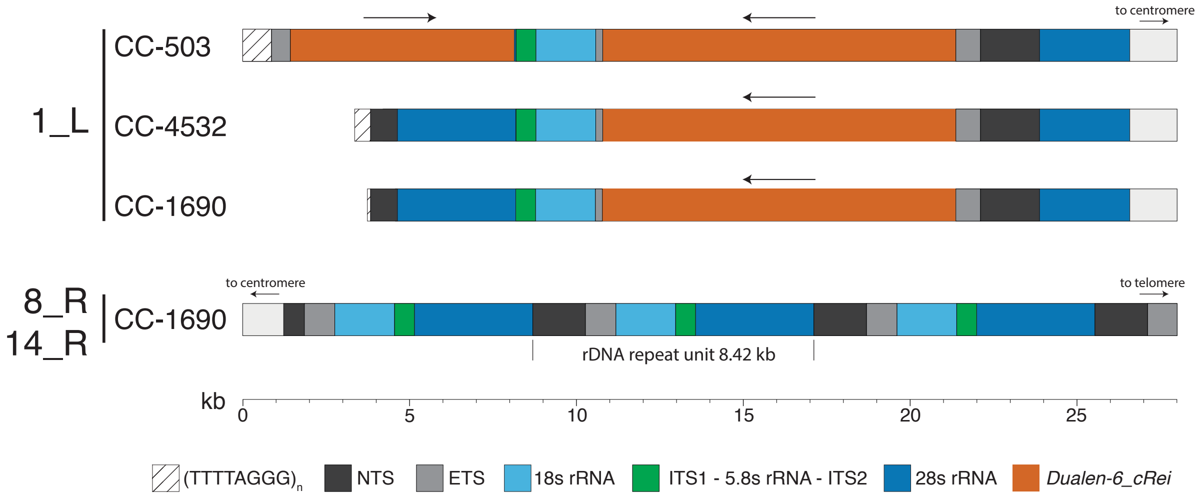


Figure S3. The left arm terminus of chromosome 1.
The rDNA array thought to be present at the left arm terminus of chromosome 1 (i.e. 1_L) is in fact incomplete. In CC-4532 v6 and CC-1690 there is one partial copy (consisting of an incomplete 28s rRNA gene) and one full rDNA copy, which is disrupted by a LINE retrotransposon (*Dualen-6_cRei*, arrows represent 3' to 5'). In CC-503 v6 part of the *Dualen* element and the ETS appear to have been duplicated and inverted, removing the full-length copy of the 28s rRNA gene. The rDNA arrays present on the right arm termini of chromosome 8 (8_R) and 14 (14_R) are shown as a comparison. ITS = internal transcribed spacer, ETS = external transcribed spacer, NTS = nontranscribed spacer.

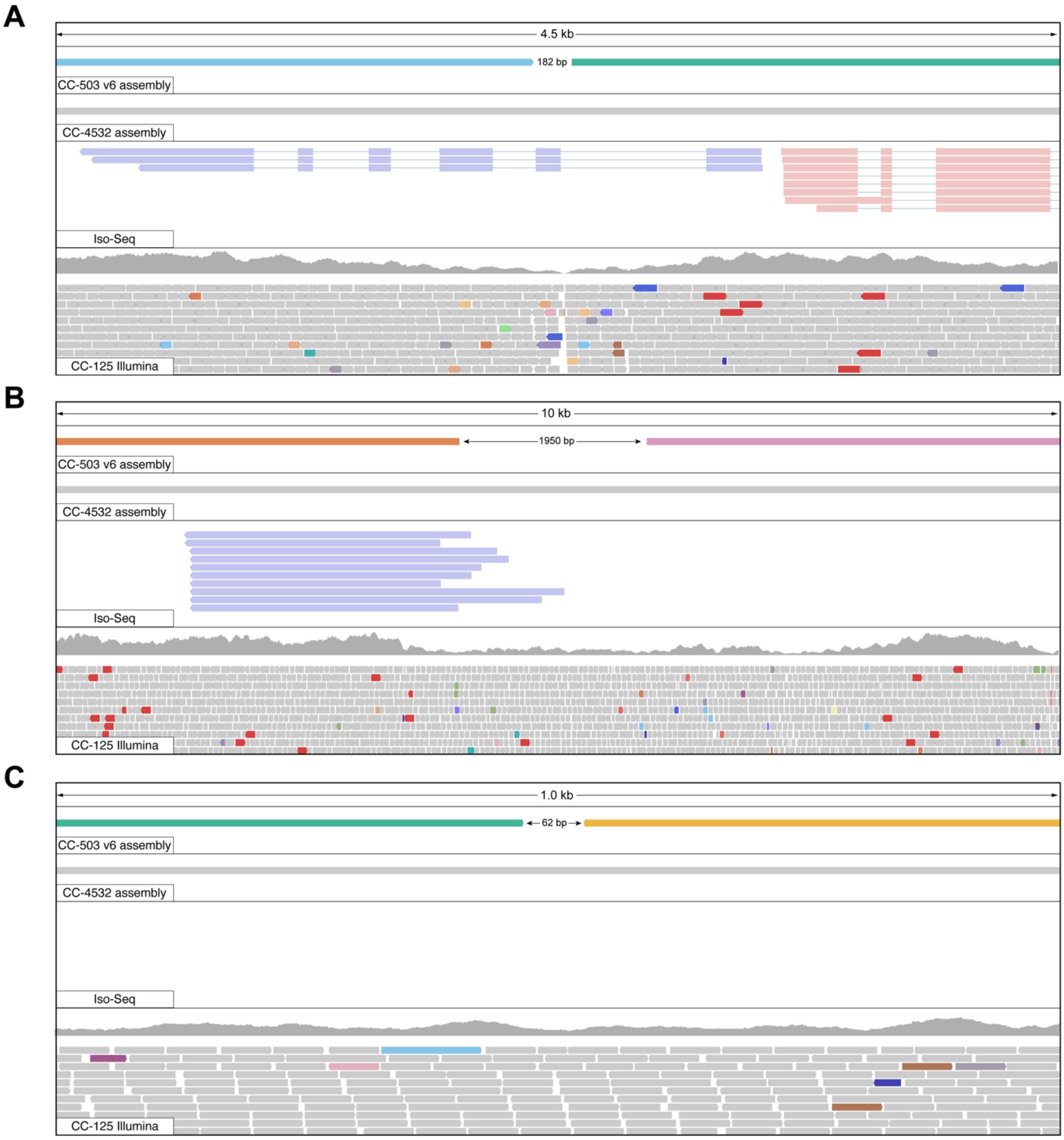


Figure S4. Browser views of DSBs associated with the reciprocal translocation shown relative to the CC-1690 assembly.

Data tracts from top to bottom are: i) alignment of CC-503 v6 assembly (colours match Figure 4), ii) CC-4532 v6 assembly, iii) Iso-Seq, iv) CC-125 re-sequencing data.

(A) DSB1, coordinates are chromosome 2: 6,577,752 - 6,582,271 bp. Note coverage drops within the middle of the deletion, however this is seen with re-sequencing data from all strains and appears to be a mapping artefact (not shown).

(B) DSB2, coordinates are chromosome 9: 2,953,140 - 2,963,530 bp.

(C) DSB3, coordinates are chromosome 2: 7,501,573 - 7,502,576 bp.

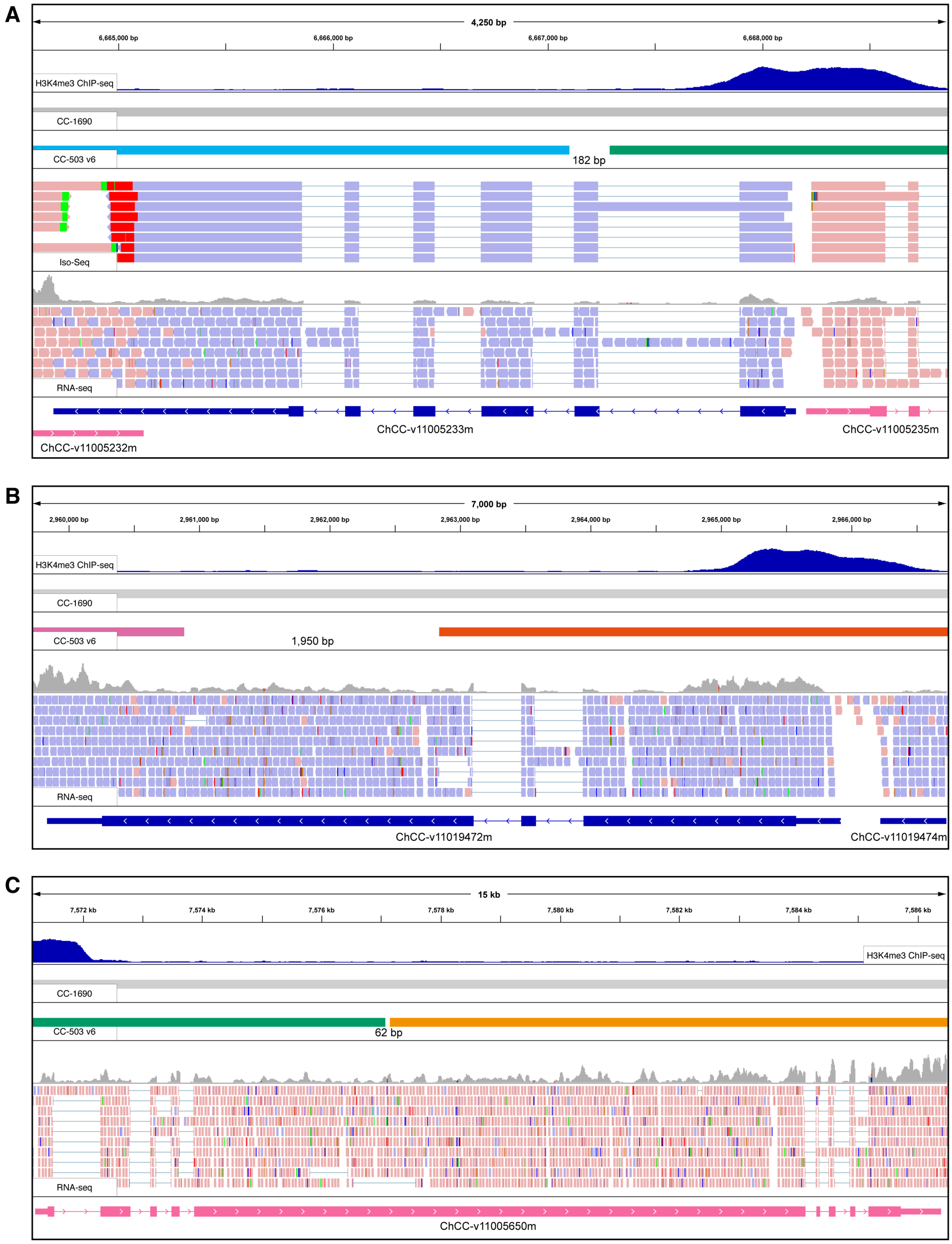


Figure S5. Browser views of DSBs associated with the reciprocal translocation shown relative to the CC-4532 v6 assembly.

Data tracts from top to bottom are: i) H3K4me3 ChiP-Seq data marking active promoters, ii), alignment of CC-1690 assembly, iii) alignment of CC-503 v6 assembly (colours match Figure 4), iv) Iso-Seq (only **A**, no full-length reads for **B**, **C**), v) RNA-seq, vi) CC-4532 v6.1 gene models.

(A) DSB1, coordinates are chromosome 2: 6,664,000 - 6,668,876 bp.

(B) DSB2, coordinates are chromosome 9: 2,959,720 - 2,966,763 bp.

(C) DSB3, coordinates are chromosome 2: 7,571,150 - 7,586,550 bp.

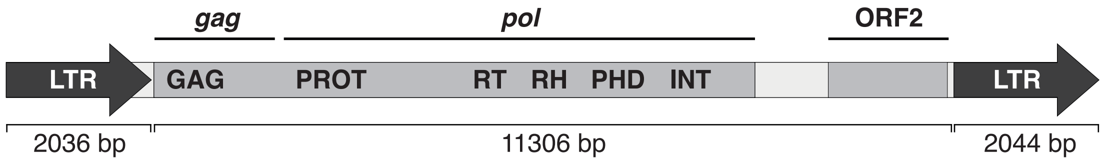


Figure S6. Structure of the *Gypsy-7a_cRei* LTR element.
LTRs are shown as block arrows, and the left LTR is missing the final 8 bp of the right LTR. The two ORFs are highlighted within the 11.3 kb internal section and the *gag* and *pol* sections of the polyprotein are highlighted. Text within ORFs show protein domains: GAG = group-specific antigen, PROT = pepsin-like aspartate protease, RT = reverse transcriptase, RH = RNAse H, PHD = plant homeodomain finger, INT = integrase.

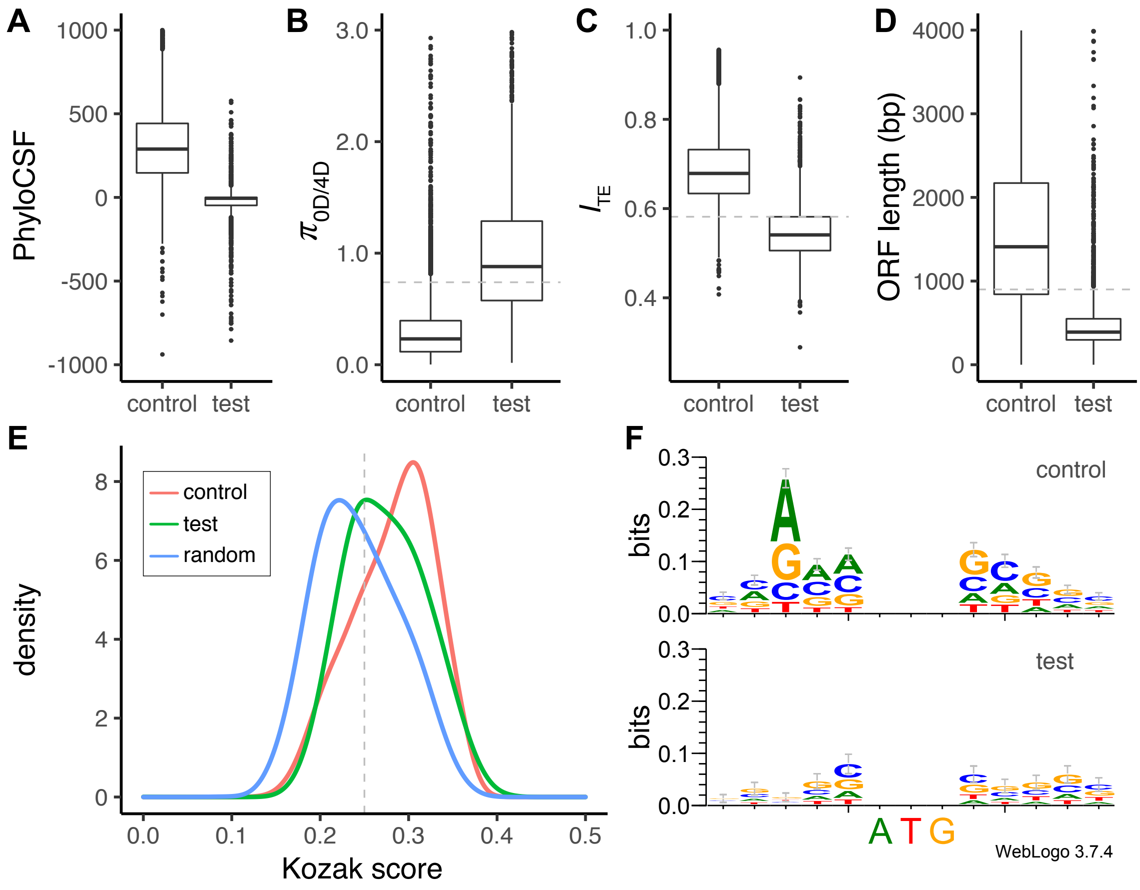


Figure S7. Coding potential analyses for CC-503 v6.1.
(A) PhyloCSF scores for control (algal homolog or functional domain, N = 15,159) and test (no homolog/domain, N = 2,418) set genes. Scores were calculated based on 8-species whole-genome alignment. Test set genes scoring <1 failed (N = 2,022).
(B) Ratio of genetic diversity at zero-fold and four-fold degenerate sites, based on whole-genome re-sequencing data of 17 *C. reinhardtii* Quebec field isolates. Test set genes with ratios >0.739 (the 95th percentile of control set) failed (N = 1,760).
(C) Codon usage as quantified by the index of translation elongation. Test set genes with values <0.582 (the 5th percentile of control set) failed (N = 1,815).
(D) ORF length minus any microsatellite or satellite DNA. Test set genes with ORFs <900 bp (N = 2,185) that failed an appropriate number of other tests were included in the low coding potential set.
(E) Kozak scores for the 5 bp up and downstream of start codons, calculated by comparison to a reference Kozak sequence. Control distribution was estimated from the half of the control set genes not used to generate the reference, and the random distribution was generated from 10,000 random sequences with an expected GC content ~64%. Test set genes with scores <0.25 failed (N = 1,407).
(F) Kozak sequence logos produced using WebLogo 3 (Crooks et al. 2004). Control logo was generated from a random half of control set genes and used as the reference for generating the distributions above.

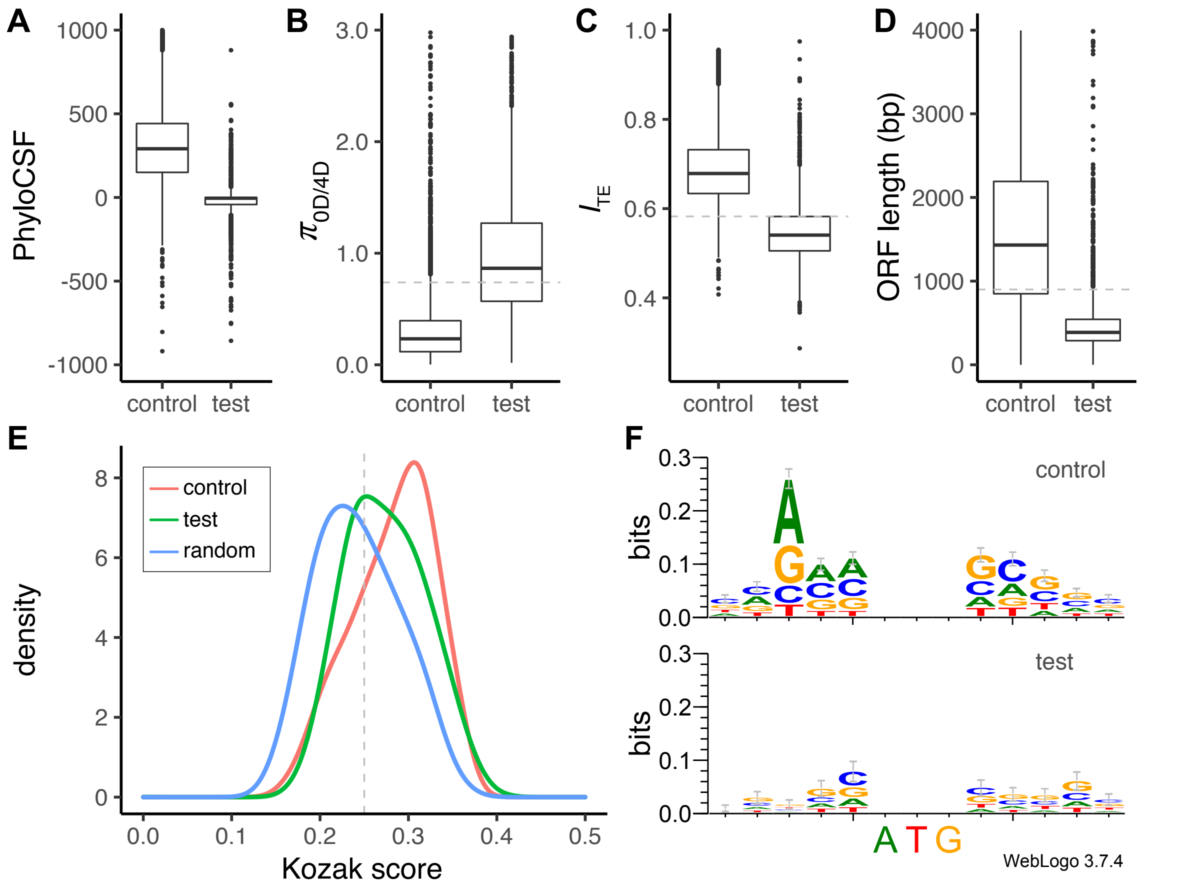


Figure S8. Coding potential analyses for CC-4532 v6.1.
(A) PhyloCSF scores for control (algal homolog or functional domain, N = 15,237) and test (no homolog/domain, N = 2,362) set genes. Scores were calculated based on 8-species whole-genome alignment. Test set genes scoring <1 failed (N = 1,982).
(B) Ratio of genetic diversity at zero-fold and four-fold degenerate sites, based on whole-genome re-sequencing data of 17 *C. reinhardtii* Quebec field isolates. Test set genes with ratios >0.738 (the 95th percentile of control set) failed (N = 1,708).
(C) Codon usage as quantified by the index of translation elongation. Test set genes with values <0.582 (the 5th percentile of control set) failed (N = 1,776).
(D) ORF length minus any microsatellite or satellite DNA. Test set genes with ORFs <900 bp (N = 2,154) that failed an appropriate number of other tests were included in the low coding potential set.
(E) Kozak scores for the 5 bp up and downstream of start codons, calculated by comparison to a reference Kozak sequence. Control distribution was estimated from the half of the control set genes not used to generate the reference, and the random distribution was generated from 10,000 random sequences with an expected GC content ~64%. Test set genes with scores <0.25 failed (N = 1,379).
(F) Kozak sequence logos produced using WebLogo 3 (Crooks et al. 2004). Control logo was generated from a random half of control set genes and used as the reference for generating the distributions above.

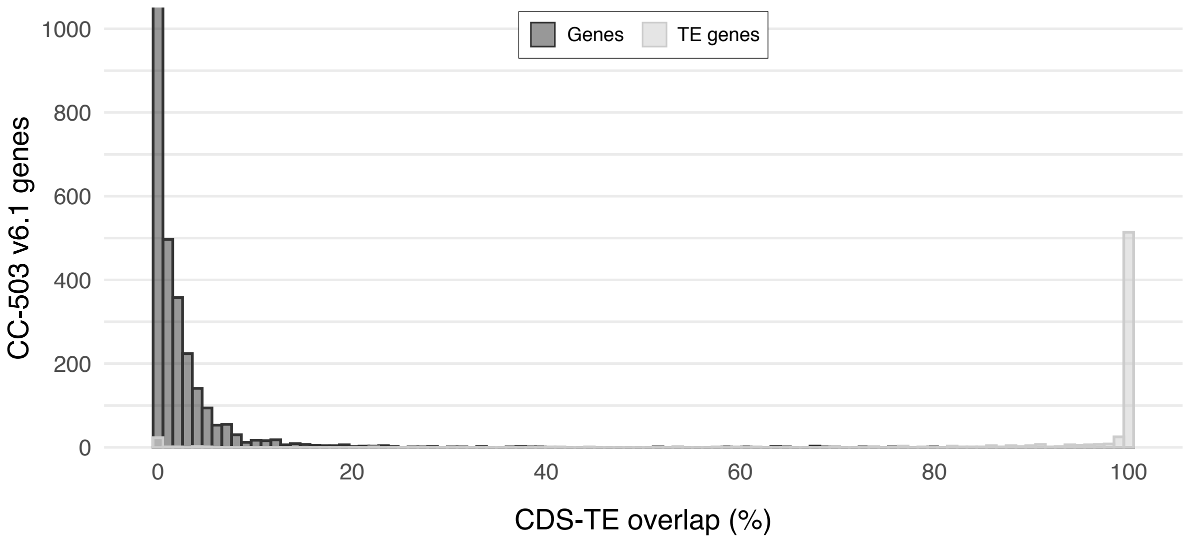


Figure S9. Intersect between CDS and TE sequence for CC-503 v6.1 standard genes and TE genes.

Appendix E

Supplementary Material for Chapter 5

The following supplementary datasets are available from the the Edinburgh Datashare repository with doi: https://doi.org/10.7488/ds/3103

Dataset S1: Annotation notes for the updated transposable element library for *Chlamydomonas reinhardtii*.

Dataset S2: *Chlamys* annotation notes.

Dataset S3: *Naiad* annotation notes.

Dataset S4. *Chlamys*-like annotation notes.

Dataset S5. *Hydra* annotation notes.

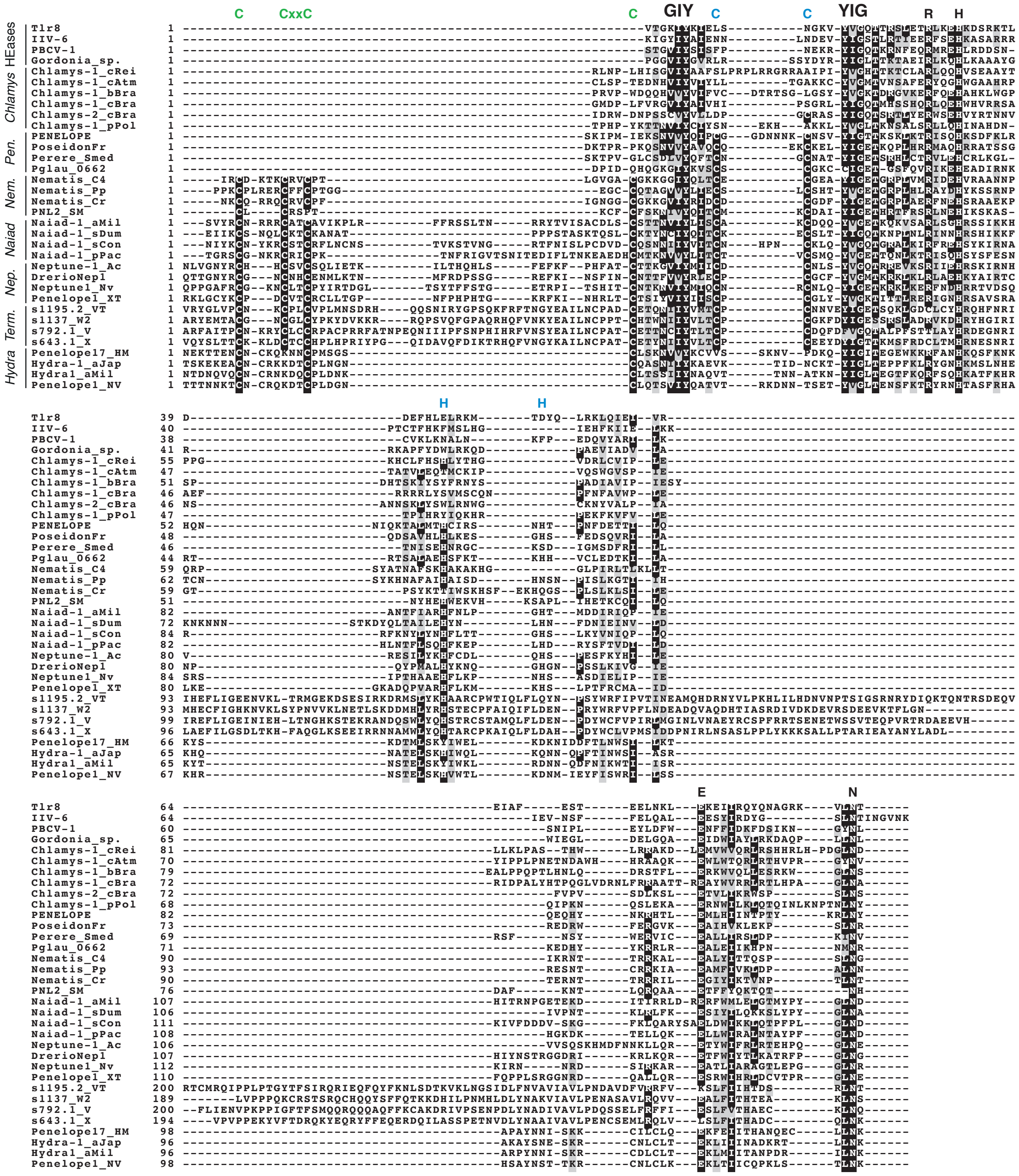


Figure S1. Alignment of the GIY-YIG EN domain from PLEs and HE_Tlr8p_PBC-V_like homing endonucleases (HEases).

Pen. = *Penelope/Poseidon*, *Nem.* = *Nematis*, *Nep.* = *Neptune*, *Term.* = *Terminon*. EN domains from Terminons were included, which are giant *Athena* elements that have acquired additional ORFs (Arkhipova et al. 2017). The CX$_{2-5}$CxxC (green) and CCHH (blue) motifs are highlighted. Alignment was produced with PROMALS3D (Pei and Grishin 2014).

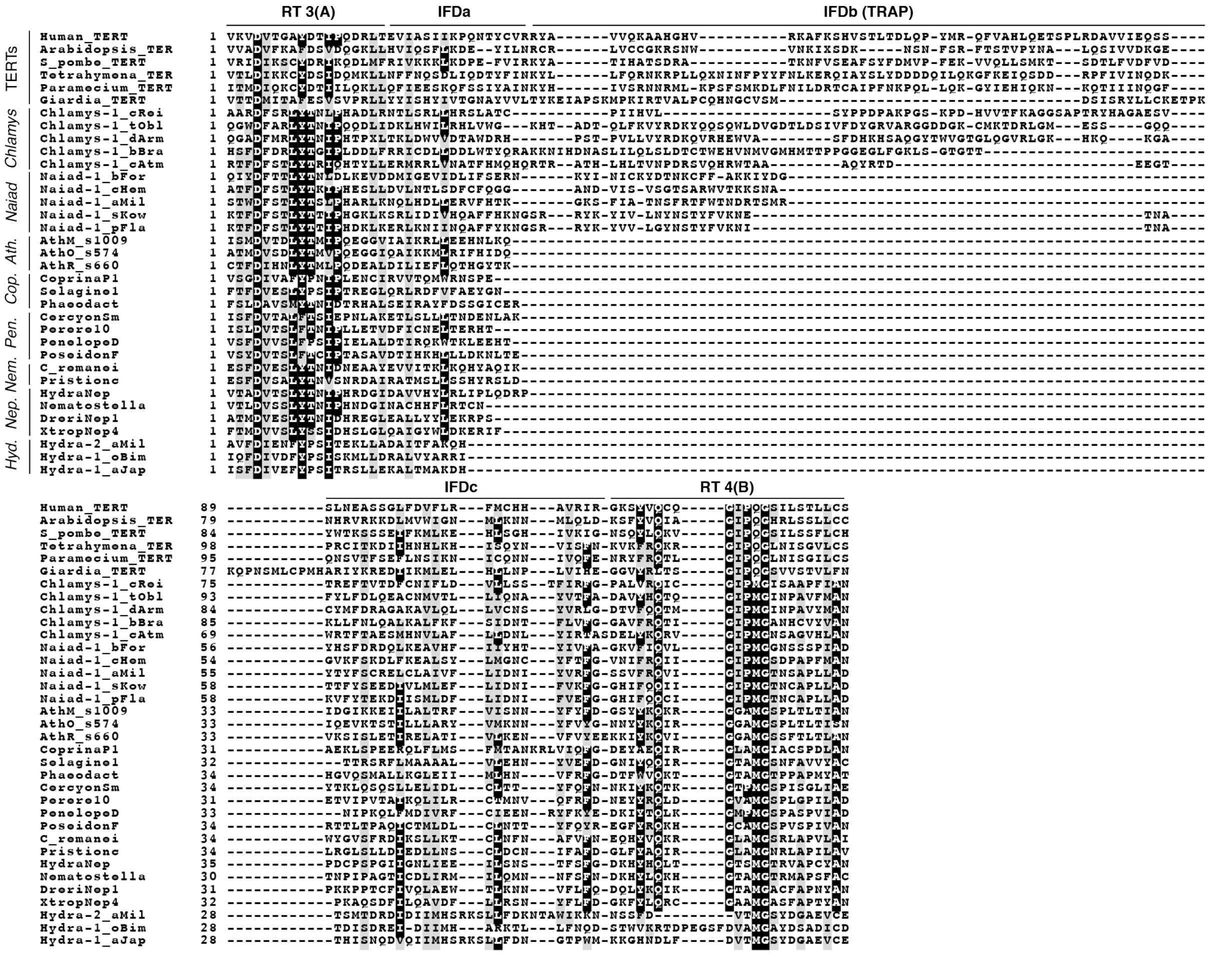


Figure S2. Alignment of RT3, IFD and RT4 regions of PLEs and TERTs.
TERT = telomerase reverse transcriptase, *Ath. = Athena*, *Cop. = Coprina*, *Pen. = Penelope/Poseidon*, *Nem. = Nematis*, *Nep. = Neptune*. Alignment was produced with PROMALS3D (Pei and Grishin 2014).

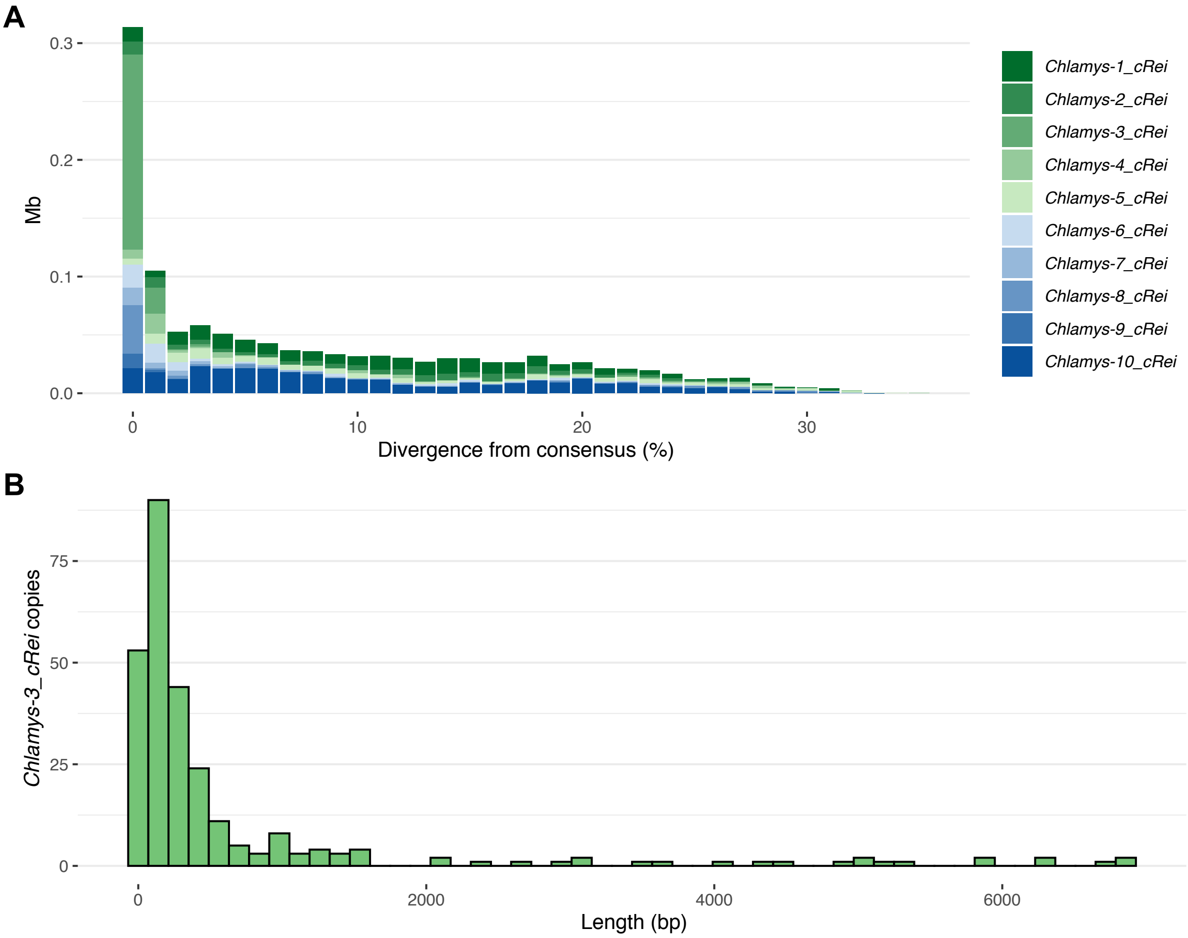
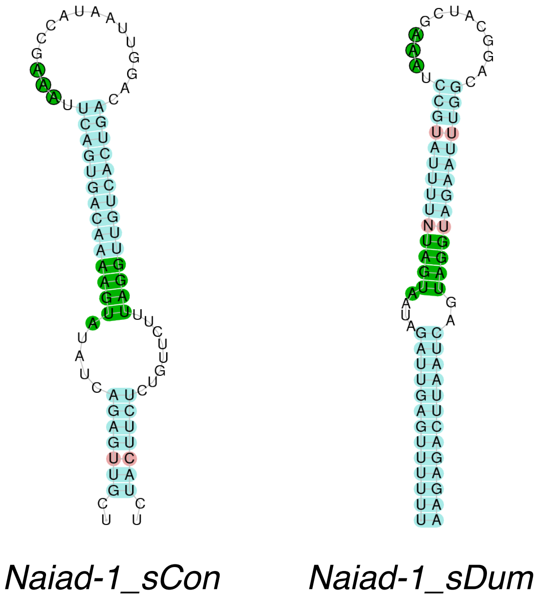


Figure S3. *Chlamys* elements in the *Chlamydomonas reinhardtii* CC-1690 assembly.
(A) Abundance vs divergence landscape plot for individual *Chlamys* families.
(B) Histogram of *Chlamys-3_cRei* copy lengths.



Naiad-1_sCon

Naiad-1_sDum

Figure S4. SECIS motifs present in 3’ of *Naiad* elements.
SECIS elements were identified by SECISearch3 (Mariotti et al. 2013) and were both designated “grade A”.

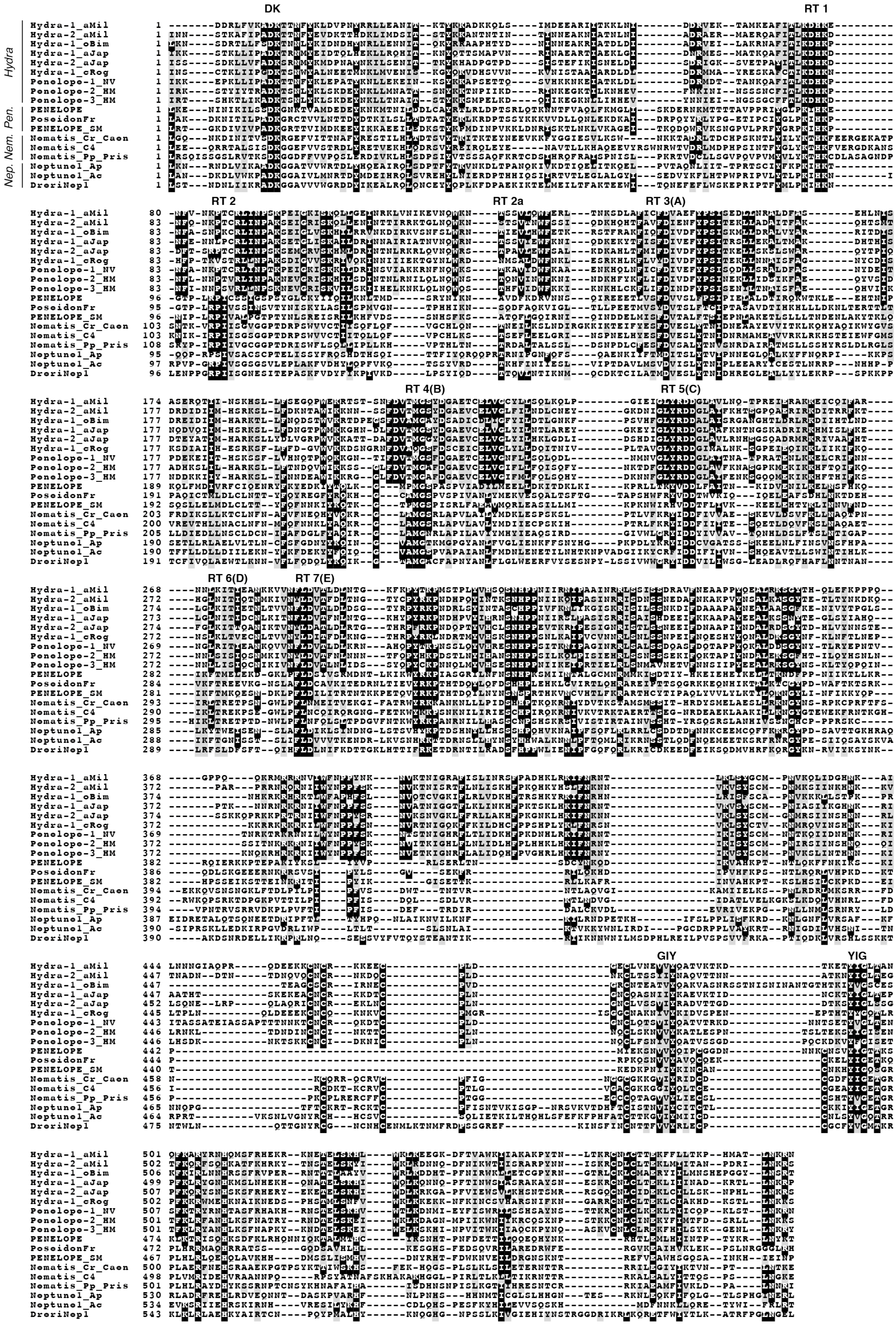

DK
RT 1
RT 2
RT 2a
RT 3(A)
RT 4(B)
RT 5(C)
RT 6(D)
RT 7(E)
GIY
YIG
Hydra
Pen.
Nem.
Nep.
Hydra-1_aMil
Hydra-2_aMil
Hydra-1_oBim
Hydra-1_aJap
Hydra-2_aJap
Hydra-1_cRog
Penelope-1_NV
Penelope-2_HM
Penelope-3_HM
PENELOPE
PoseidonFr
PENELOPE_SM
Nematis_Cr_Caen
Nematis_C4
Nematis_Pp_Pris
Neptune1_Ap
Neptune1_Ac
DreriNep1

Figure S5. Alignment of *Hydra* and selected canonical PLEs.
Pen. = *Penelope/Poseidon*, *Nem.* = *Nematis*, *Nep.* = *Neptune*. Alignment was produced with PROMALS3D (Pei and Grishin 2014).

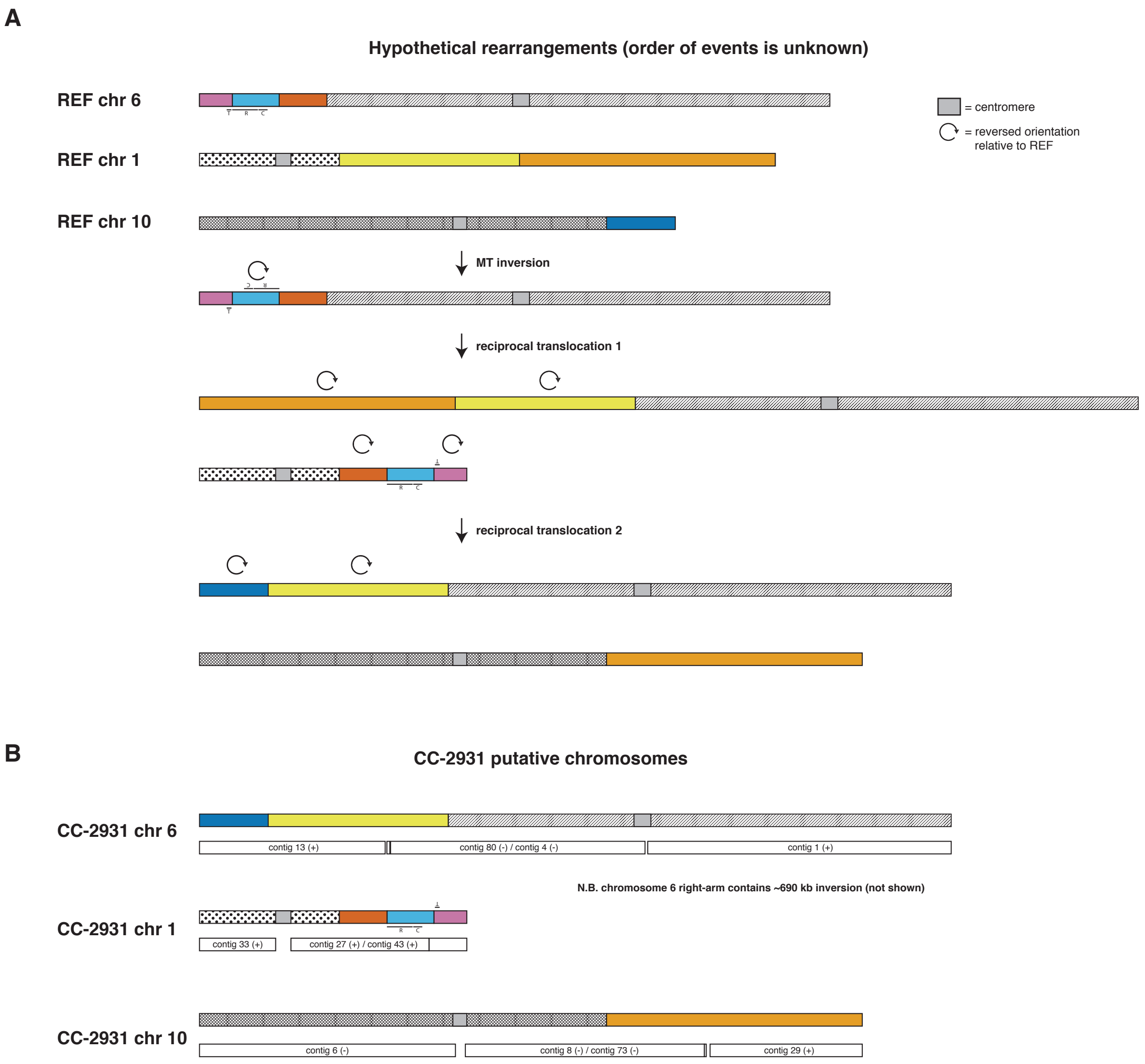

A
Hypothetical rearrangements (order of events is unknown)
REF chr 6
T
R
C
REF chr 1
REF chr 10
= centromere
= reversed orientation relative to REF
MT inversion
reciprocal translocation 1
reciprocal translocation 2
B
CC-2931 putative chromosomes
CC-2931 chr 6
contig 13 (+)
contig 80 (-) / contig 4 (-)
contig 1 (+)
N.B. chromosome 6 right-arm contains ~690 kb inversion (not shown)
CC-2931 chr 1
R
C
contig 33 (+)
contig 27 (+) / contig 43 (+)
CC-2931 chr 10
contig 6 (-)
contig 8 (-) / contig 73 (-)
contig 29 (+)

Figure S6. Putative rearrangements and translocations in the CC-2931 genome assembly.
(A) Depiction of chromosomes in standard laboratory strains (i.e. CC-1690 and CC-4532), black and white patterns or colours mark sequence blocks. The order of events is unknown, however an inversion of the light blue block within the mating type locus on chromosome 6 is implied (T, R and C domains are highlighted). A reciprocal translocation between chromosomes 1 and 6 is shown, with the yellow + orange block from the right arm of chromosome 1 exchanged with the vermillion + light blue + purple block from the left arm chromosome 6. A second reciprocal translocation is shown, with the orange block (now on the left arm of chromosome 6) exchanged with the dark blue block from the right arm of chromosome 10.
(B) Putative chromosomal assembly for chromosomes 1, 6 and 10 in CC-2931. The underlying contigs are shown as white blocks. Paired contigs (e.g. contig 4 + contig 80) were manually inferred by examining contig breaks and PacBio read alignments.