

https://dspace.org/

The molecular basis of muticellularity

Featherston, Jonathan

2018
https://hdl.handle.net/10539/26677
Featherston, Jonathan (2018) The molecular basis for multicellularity, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/26677>

Downloaded from a DSpace repository. https://dspace.org/

The Molecular Basis Of Multicellularity

Jonathan Featherston

A Thesis submitted to the Faculty of Health Sciences

University of Witwatersrand

Doctor of Philosophy

Johannesburg

2018

Declaration

I, Jonathan Featherston, declare that this thesis is my own work. It is being submitted for the degree of Doctor of Philosophy at the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university

……

-------- day of ……., 2018

“Out, out, brief candle! Life’s but a walking shadow, a poor player that struts and frets his hour upon the stage and is heard no more. It is a tale told by an idiot, full of sound and fury, signifying nothing”. “False face must hide what the false heart doth know”.

From Macbeth

“A life of science is a life of endless labour, without any reward other than the pure exalted enjoyment which it brings to those minds who pursue it from a real love of its superior character"

Burchell

To my family. All my love.

Publications and Presentations related to this thesis:

Publications

1. Takashi Hamaji, Bradley J.S.C Olson, Jonathan Featherston, Pierre M Durand, Patrick Ferris, Richard E Michod, Hisayoshi Nozaki, David R Smith and Patrick J Keeling. (2013). Organelle genome complexity scales positively with organism size in volvocine green algae. *Molecular Biology and Evolution.* 30(4):793-797.

2. Michael Del Vasto, Francisco Figueroa-Martinez, Marilea Gonzalez, Adrian Reyes-prieto, Jonathan Featherston, Pierre M Durand and David Roy Smith. (2015). Massive and widespread organelle genomic expansion in the green algal genus Dunaliella*. Genome Biology and Evolution*. 7(3):656-663.

3. Jonathan Featherston, Yoko Arakaki, Hisayoshi Nozaki, Pierre M Durand and David R Smith (2016). Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in the volvocine green algae. *European Journal of Phycology*. 51(4):369-377

4. Yoko Arakaki, Takayuki Fujiwara, Hiroko Kawai-Toyooka, Jonathan Featherston, Pierre Durand, Shin-ya Miyagishima, Hosayoshi Nozaki. (2017). Evolution of cytokinesis-related protein localization during the emergence of multicellularity. *BMC Evolutionary Biology*. 17:243

5. Jonathan Featherston, Yoko Arakaki, Erik Hanschen, Patrick Ferris, Rick Michod, Bradley Olson, Hisayoshi Nozaki, Pierre Durand. The 4-celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocines lineage. (2018). *Molecular Biology and Evolution*. 35(4):837-854.

Conference Presentations

1. Jonathan Featherston and Pierre Durand. (2012). Genome Sequencing of Volvocine Algae. South African Society for Biochemistry and Molecular Biology, Drakensburg, South Africa.

2. Jonathan Featherston and Pierre Durand (2015). The Draft Genome Sequence of *Tetrabaena socialis*. Third International Volvox Conference, Cambridge, UK.

Additional publications during the course of the degree, unrelated to the primary aims and objectives

1. Amanda Dodd, Jonathan Featherston, Dirk Swanevelder and Karl Rumbolt. (2013). Draft Genome Sequencing of *Streptomyces albulus* Strain CCRC 11814, an E-Poly-L-Lysine-Porducting Actinomycete. *ASM:Genome Announcements*. 1(5).

2. Tisetso E. Lephoto, Jonathan Featherston and Vincent M Gray. (2015). Draft Whole-Genome Sequence of *Serratia sp*. Strain TEL, Associated with *Oscheius sp* TEL-2014 (Nematoda:Rhabditidae) Isolated from Grassland in South Africa. *ASM: Genome announcements*. 3(4):e000747-15.

3. Stephanie Naidoo, Jonathan Featherston and Vincent M Gray. (2015). Draft Whole-Genome Sequence and Annotation of the Entomopathogenic Bacterium *Xenorhabdus khoisanae* Strain MCB. *ASM: Genome announcements*. 3(4):e00872-15.

4. Boipelo Mothupi, Jonathan Featherston and Vincent M Gray (2015). Draft-Whole Genome Sequence and Annotation of *Xenorhabdus griffiniae* Strain BMMCB Associated with the South African Entomopathegenic Nematode *Steinernema Khoisanae* strain BMMCB. *ASM: Genome announcements*. 3(4):e00785-15.

5. Stephanie Naidoo, Boipelo Mothupi, Jonathan Featherston and Vincent M Gray. (2015). Draft Genome Sequence and Assembly of *Photorhabdus heterorhabditis* Strain VMG, a Bacterial Symbiont Associated With The Enteropathogenic Nematode *Heterorrhabditis zealandica*. *ASM: Genome announcements*. 3(5):e01279-15.

6. Lee-Anne O Soobramoney, Jonathan Featherston and Vincent M Gray. Draft Whole-Genome Sequence of *Xenorhabdus sp*. Strain GDC328, Isolated from the Indigenous South African Nematode Host

Steinernema khoisanae. *ASM:Genome announcements*. 3(5):e01239-15.

7. Ben Mans, Jonathan Featherston, Minique de Castro and Ronel Pienaar (2017). Gene duplication and protein evolution in tick-host interactions. *Frontiers in Cellular and Infection Microbiology*. 7:413

8. Enzo Klein, Jonathan Featherston, Jasper Rees and Thomas Debener. (2017). A draft genome sequence of the black spot fungus of roses reveals a high degree of gene duplications. *PLoS One*. 12(10):e0185310

ABSTRACT

The volvocine lineage of chlorophytes is a model lineage for investigating the evolution of multicellularity, sexual dimorphism and other biological processes. Extant members of the lineage include unicellular to many celled multicellular taxa with germ-soma division of labour. Investigating the genome sequences of modern volvocines furthers our understanding of the molecular basis of multicellularity. Amongst the simplest of multicellular organisms in the lineage is the homothallic 4-celled *Tetrabaena socialis*, which makes it ideal for investigating the genomic basis for the initial transition to simple multicellularity. With the aim of exploring the molecular basis for the evolution of multicellularity *de novo* nuclear and organelle genome assemblies as well as a *de novo* transcriptome assembly of *Tetrabaena socialis* NIES-571 were generated. Comparative genomic analyses with available volvocine genomes were performed. The organelle genomes of *Tetrabaena socialis* were found to be highly inflated with non-coding DNA and the mitochondrial genome is circular indicating that a circular mitochondrial genome was present near the origin of multicellularity. Inflated organelle genomes are presumed be due to population genetic effects and neutral evolution. As per previous analyses of volvocine genomes comparative genomic analyses undertaken in this thesis demonstrated that the evolution of multicellularity was not associated with extensive gains in genome or proteome complexity.

Various analyses identified modifications to the ubiquitin proteasomal pathway (UPP) that were associated with the origin of multicellularity. Targeted ubiquitin-mediated degradation of cell cycle effector molecules is a potential mechanism for regulation of cell division number in the volvocine lineage. Gene families gained at the origin of multicellularity were identified and found to be enriched in developmental genes and lineage specific genes. Few genes gained are known only from multicellular organisms and, therefore, the analysis of gene family gain and loss does not support shared genetic mechanisms for multicellularity but instead that the evolution of multicellularity involved lineage specific genetic mechanisms. Mating-locus genes that in heterothallic volvocines are limited to either gender, were identified in the genome of the homothallic *Tetrabaena socialis* (the first homothallic taxon in the lineage to be sequenced). The presence of gender-limited mating-locus genes in the genome of *Tetrabaena socialis* demonstrates how a single strain can produce both *plus* and *minus* mating-types. The genome sequence of *Tetrabaena socialis* adds to a growing collection of volvocine genomes for analysis and as one of the simplest in the lineage it is of unique significance.

ACKNOWLEDGEMENTS

1. I would like to thank my supervisor Dr Pierre Durand for all his support, mentorship and guidance. You have taught me **how** to think.
2. I would like to thank my family for their love, support, motivation and occasional chastisement.
3. I would like to thank the University of Witwatersrand Faculty of Health Science for a Faculty research award.
4. I would like to thank Dr Yoko Arakaki for her contribution to the publication of the *Tetrabaena socialis* genome and for the provision of nucleic acids for genome and transcriptome sequencing.
5. I would like to thank members of the Michod laboratory, including Professor Rick Michod, at the University of Arizona for their contributions to the publication of the *Tetrabaena socialis* nuclear genome.
6. I would like to thank Dr David Smith for his insights and assistance with the publication of the *Tetrabaena socialis* organelle genomes.
7. I would like to thank Dr Jasper Rees and the Agricultural Research Council for access to their facilities.

TABLE OF CONTENTS

DECLARATION..... ii

DEDICATION..... iv

PUBLICATIONS AND PRESENTATIONS RELATED TO THESIS..... v

ABSTRACT..... ix

ACKNOWLEDGEMENTS..... xi

TABLE OF CONTENTS xii

LIST OF FIGURES..... xix

LIST OF TABLES xxii

LIST OF ABBREVIATIONS xxiii

LIST OF SYMBOLS xxv

GLOSSARY OF TERMS..... xxvi

1. INTRODUCTION AND LITERATURE REVIEW..... 1

1.1 The Evolution of Complexity..... 6

1.2 The Evolution of Multicellularity..... 10

1.3 The Volvocine Lineage..... 16

1.3.1 Brief Introduction to Volvocine Life Cycles..... 22

1.3.2 Kirk's 12 Step to Multicellularity..... 27

1.3.2.1 Incomplete Cytokinesis..... 29

1.3.2.2 Incomplete Inversion of the Embryo..... 32

1.3.2.3 Rotation of Basal Bodies..... 30

1.3.2.4 Establishment of Organismal Polarity..... 30

1.3.2.5 Transformation of Cell Walls into an Extracellular Matrix (ECM)..... 31

1.3.2.6 Genetic Modulation of Cell Number..... 32

1.3.2.7 Complete Inversion of the Embryo..... 32

1.3.2.8 Increased Volume of the ECM…… 34
1.3.2.9 Partial Division of Labour…… 34
1.3.2.10 Complete Germ and Soma Division of Labour……..…… 34
1.3.2.11 Asymmetric Division…… 35
1.3.2.12 Bifurcation of the Cell Division Programme……..….. 35
1.3.3 Driving Forces Behind Increasing Complexity in the Volvocine Algae…38
1.3.4 Genetic Developments within the Volvocine Lineage.…….…. 41
1.3.4.1 Incomplete Cytokinesis…… 41
1.3.4.2 Inversion……..…… 43
1.3.4.3 Basal Body Rotation…… 44
1.3.4.4 Organismal Polarity…….. 44
1.3.4.5 Transformation of the Cell wall into an ECM and the Expansion of the ECM……..…… 45
1.3.4.6 Genetic Modulation of Cell Number and the Modification of the Cell Cycle……..…… 48
1.3.4.6.1 The retinoblastoma Homologue…… 51
1.3.4.6.2 Cyclin Dependent kinase G1…… 53
1.3.4.6.3 Other Cell Cycle Determining Genes…… 54
1.3.4.7 The Molecular Underpinnings of Germ and Soma Differentiation……... 58
1.3.5 Genomic Developments in the Volvocine Lineage……... 64
1.3.5.1 Repetitive Elements and Synteny…….. 66
1.3.6 Sexual Reproduction of The Mating Locus…… 72
1.3.7 Volvocine Organelle Genomes…… 76
1.3.8 *Tetrabaena Socialis*…… 77
1.3.9 A Model for the Evolution of *Tetrabaena Socialis*…….. 79
1.4 Conclusions…….. 81

2. ORGANELLE GENOME EVOLUTION IN THE VOLVOCINES… 82

2.1 Summary and Background of Manuscripts…….. 82

3. THE NUCLEAR GENOME OF TETRABAENA SOCIALIS…..... 85

3.1 Summary of Manuscripts…… 85

4. THE TETRABAENA SOCIALIS MATING-LOCUS……...…. 88

4.1 Abstract…….. 88

4.2 Introduction…… 88

4.2.1 Volvocine Sexual Reproduction…… 88

4.2.1.1 Volvocine Mating Loci…….. 91

4.2.1.2 Gender Associated and Sex determining Genes…… 93

4.2.1.3 Homothallism and Heterothallism……. 95

4.3 Methods…… 98

4.3.1 Identification of Mating-Locus Genes and Scaffolds……..... 98

4.3.2 Synteny……..…….…. 98

4.3.3 Polymerase Chain reaction Confirmation of *MID*, *FUS1* and *MTD1* Gene Structures……..... 98

4.3.4 Phylogenetic Analyses…….. 99

4.4 Results…… 100

4.4.1 Scaffolds Housing Mating Locus Genes…….. 100

4.4.2 Putative Mating Locus Genes…… 102

4.4.2.3 Gender-Limited Genes……..…… 105

4.5 Discussion……... 110

4.5.1 GC Content…… 111

4.5.2 Homothallism in the Volvocine Lineage…… 112

4.5.2.1 Pseudo-Mating Types…… 114

4.5.3 Evolutionary Implications of the Transition to Homothallic Reproduction..... 115

5. FINAL THOUGHTS AND CONCLUSIONS..... 121

5.1 General..... 126

5.2 Future Recommendations..... 129

6. APPENDICES..... 131

6.1 Introduction to Appendices..... 131

7. APPENDIX A: GENOME AND TRANSCRIPTOME ASSEMBLY..... 132

7.1 Methods..... 132

7.1.1 Next Generation Sequencing..... 132

7.1.2 Evaluation of de novo Genome Assemblers..... 132

7.1.2.1 Scaffolders..... 135

7.1.2.2 Post-assembly..... 136

7.1.2.3 Basic Assembly Statistics..... 137

7.1.2.4 Coding-content of Genome Assemblies..... 138

7.1.3 Genome Cleaning..... 139

7.1.4 Transcriptome Assembly..... 139

7.1.4.1 Data Trimming..... 139

7.1.4.2 Trinity Assembly..... 140

7.1.4.3 Bridger de novo Assembly..... 140

7.1.4.4 Transcriptome Evaluation..... 140

7.2 Results..... 141

7.2.1 Sequencing Output..... 141

7.2.2 Testing Assemblers..... 142

7.2.3 Final Assembly…… 145
7.2.4 Kmer Analysis…… 150
7.2.5 Genome Coding Content…… 155
7.2.6 Genome Contamination…… 158
7.2.7 De Novo Transcriptome…… 159
7.2.7.1 Trimming…… 159
7.2.7.2 Transcriptome Assembly Results…… 160
7.3 Discussion…… 162
7.3.1 Nucleur Genome Quality…… 162
7.3.2 Transcriptome…… 167
7.3.2.1 General…… 167
7.3.2.2 Quality Assessment of Bridger De Novo 25mer Assembly…… 168

8. APPENDIX B: NUCLEAR GENOME ANNOTATION…… 170
8.1 Methods…… 170
8.1.1 Annotation of Gene models…… 170
8.1.2 Functional Annotations…… 171
8.2 Results…… 173
8.2.1 Gene Models…… 173
8.2.2 Functional annotation…… 177
8.3 Discussion…… 179
8.3.1 Gene Models and Summary Statistics. …… 179

9. APPENDIX C: ADDITIONAL COMPARATIVE GENOMIC ANALYSES…… 182
9.1 Methods…… 182
9.1.1 Syntenic re-arrangements…… 182

9.1.2 Phylogenomics…… 182
9.1.3 Comparative genomics: general functional annotations…… 184
9.1.3.1 Differences in general functional annotations……... 184
9.1.3.2 Mercator and Gene Ontology term analysis……... 184
9.1.3.3 Volvocine genes of interest……... 185
9.1.3.3.1 Prochnik table genes…… 185
9.1.3.3.2 Inversion A (inva/iar1)…… 186
9.1.3.3.3 Cyclins…….…….. 188
9.1.3.3.4 Retinoblastoma Gene…….. 190
9.1.3.3.5 Fasciclin domain-containing proteins……. 190
9.2 Results…… 192
9.2.1 Synteny…….…….. 192
9.2.2 Phylogenomic analysis of Volvocine Taxonomy…….. 194
9.2.3 Comparative Genomics: Global annotations associated with the Evolution of Multicellularity…….. 197
9.2.3.1 Mercator assigned MapMan annotations…… 197
9.2.3.2 Gene Ontology Analysis…… 200
9.2.3.3 Specific Gene Families…….. 201
9.2.3.4 The Retinoblastoma Homologue…….. 207
9.2.3.5 Fasciclin-domain Containing Proteins…… 211
9.3 Discussion……... 214
9.3.1 Syntenic rearrangements…… 214
9.3.2 Phylogeny…….. 215
9.3.3 High Level Descriptors…… 216
9.3.4 Fasciclin domain containing proteins…….…… 217
10. APPENDIX D: GENOMIC INSIGHTS OF KIRK'S 12 STEPS.... 219
10.1 Incomplete Cytokinesis…… 219
10.2 Partial and Complete Inversion…… 223

10.3 Rotation of Basal Bodies……... 226

10.4 Establishment of Organismal Polarity……. 228

10.5 Transformation of the Cell wall into an ECM and the Expansion of the ECM……..… 231

10.6 Genetic Control of Cell number…… 232

10.6.1 Cyclins…… 233

10.6.2 Retinoblastoma Gene……..… 234

10.6.3 Additional regulators of the cell cycle…… 236

10.7 Germ-soma division of labour, Asymmetric Division and the Bifurcation of the Cell Division Program……. 238

11. APPENDIX E: TURNITIN REPORT……. 240

12. APPENDIX F: ETHICS WAIVER……..……...255

REFERENCES…… 256

List of Figures

Figure 1.1: A recent phylogenetic construction of volvocine relationships……..Pg:19

Figure 1.2: The respective life cycles of the unicellular Chlamydomonas reinhardtii and Volvox carteri…….Pg:26

Figure 1.3: Ancestral character states of Kirk's 12 steps plotted against a volvocine phylogeny……..Pg:28

Figure 1.4: Central to Peripheral as well as anterior to posterior organismal Polarity…….Pg:31

Figure 1.5: Electron micrograph images of the developing Volvox embryo undergoing inversion…….Pg:33

Figure 1.6: A working model of the Genetic underpinnings of Volvox germ-soma differentiation…….....Pg:61

Figure 1.7: Synteny between Chlamydomonas, Gonium and Volvox……...Pg:71

Figure 1.8: Types of sexual reproduction in the volvocine lineage……....Pg:73

Figure 1.9: A light microscope image of *Tetrabaena socialis*…….....Pg:79

Figure 4.1: Mating locus synteny……...Pg:102

Figure 4.2: Mid-rooted maximum likelihood phylogeny of MID proteins…….Pg:106

Figure 4.3: MAFFT alignment of HAP2/GSC1 proteins from *Chlamydomonas reinhardtii*, *Tetrabaena socialis*, *Gonium pectorale* and *Volvox carteri*.....Pg:109

Figure 7.1: A cumulative histogram of scaffolds (blue) and contigs (red) for the final assembly of *Tetrabaena socialis*…….Pg:149

Figure 7.2: NxG diagram of scaffold and contig length (y-axis) versus percentage of genome (x-axis)…….....Pg:150

Figure 7.3: Kmer distribution plot generated by Allpaths-LG……Pg:152

Figure 7.4: Cumulative histogram of 25mer kmers……..……..Pg:154

Figure 8.1: A screenshot of *Tetrabaena socialis* MAKER2 gene model TSOC_008923-RA as viewed with Apollo genome browser…….Pg:172

Figure 8.2: Cumulative histogram of annotation edit distance scores…….Pg:173

Figure 8.3: Pie chart of MapMan terms assigned to *Tetrabaena socialis* Proteins……..….Pg:178

Figure 9.1: Synteny between *Chlamydomonas reinhardtii* and *Tetrabaena socialis* viewed in circular format…….Pg:193

Figure 9.2: Dot plots of synteny between *Chlamydomonas reinhardtii* and *Tetrabaena socialis*……..Pg:194

Figure 9.3: RAxML phylogenetic analysis of proteomes from genome sequenced chlorophytes with additional transcriptome derived proteomes from taxa in the Chlamydomonadales order…….Pg:196

Figure 9.4: Neighbour-joining phylogeny of InvA/IAR1 produced from an alignment of 237 amino acids……...Pg:204

Figure 9.5: Maximum likeilhood phylogeny of D-type volvocine cyclins……..Pg:206

Figure 9.6: Gene structures of various cyclin D proteins……...Pg:207

Figure 9.7: Mid-point rooted phylogeny of fasciclin-domain containing volvocine proteins……..….Pg:213

List of Tables

Table 1.1: Volvocine genome assembly statistics……..…..……Pg:66

Table 4.1: Primer sequences used to confirm the gene presence and structure for *MID*, *MTD1* and *FUS1* genes……..…….……Pg:99

Table 4.2: Genes likely or unlikely to reside within a hypothetical *Tetrabaena socialis* mating-locus based on gene location ……..…..Pg:104

Table 4.3: GC content of *MTD1*, *FUS1* and *MID* introns and exons……..Pg:107

Table 7.1: Different assemblers tested…….Pg:135

Table 7.2: Sequencing yields across various library types and platforms.……...Pg:142

Table 7.3: Results from five assemblers tested with simulated *Chlamydomonas reinhardtii* data……..……...Pg:144

Table 7.4: Results from optimised assemblies of *Tetrabaena socialis*…….……...Pg:145

Table 7.5: Testing different assembly mergers with Allpaths-LG as the master assembly……..……....Pg:147

Table 7.6: Iterative Improvements towards the final assembly……Pg:148

Table 7.7: CEGMA results for volvocine genomes as well as the genomes of *Steinernema* MCB28 and *Rosa spp*……..……..Pg:158

Table 7.8: Summary of transcriptome assembly results……..Pg:161

Table 8.1: Summary Statistics of volvocine annotations……...……Pg:175

Table 8.2: Orthologues between volvocine proteomes…….……Pg:176

Table 8.3: DOGMA assessment of proteome completeness for volvocine Algae…..……Pg:177

Table 8.4: Blast results of Tetrabaena Socialis Proteins against UNIPROT and the Proteomes of other Volvocines……...…Pg:178

Table 9.1: MapMan Terms significantly different ($p<0.05$) between the unicellular *Chlamydomonas reinhardtii* and the Multicellular Volvocines……..Pg:199

Table 9.2: Gene Ontology terms significantly different ($p<0.05$) between the Unicellular *Chlamydomonas reinhardtii* and the multicellular volvocines….....Pg:201

Table 9.3: The length of the linker region between the retinoblastoma A and B domain in Chlorophyte orthologues of MAT3/RB……Pg:209

Table 9.4: Phosphorylation targets within the linker between the retinoblastoma B and C-terminal domain in Chlorophyte orthologues of MAT3/RB….Pg:210

List of Abbreviations

AED – Annotation-edit Distance

AIC – Akaike Information Criterion

Algal-CAM – Algal Cell Adhesion Molecule

BAC – Bacterial Artificial Chromosomes

BIC – Bayesial Information Criterion

BLAST – Basic Local Alignment

CDA – Conserved Domain Arrangements

CDK - Cyclin-Dependent Protein Kinases

cpDNA – Chloroplast DNA

CTAB – Cetyl trimtheylammonium bromide

CYC – Cyclin

dN – Non-synonymous nucleotide substitutions

dS – Synonymous nucleotide substitutions

DT - Decision Theory Criterion Framework

ECM - Extracellular Matrix

ETI – Evolutionary Transition in Individuality

G1 – Gap phase in the cell cycle. Part of the interphase period

GB – Gigabases or Gigabytes

GO – Gene Ontology

HPC – High Performance Cluster Computing

HSP – High-Scoring segment Pair

KOG – Clusters of Orthologous Groups

LTR – Long Tandom Repeat

MMPs – Matrix Metalloproteases

MP – Mate-Pair

mtDNA – Mitrochondrial DNA

MTE - Major Transitions in the Evolution

MT-locus – Mating Locus

MYA – Million Years Ago

NGS – Next Generation Sequencing

NR – Nucleotide Resource Database

ORF – Open Reading Frame

ORIMC – Origin of Multicellularity

PCR – Polymerase Chain Reaction

PPX – Hidden Markov Model Protein Profile

RLS - RegA-related

RNAseq – Next generation sequencing of RNA

S – DNA synthesis phase of the cell cycle

TEM – Transition Electron Microscopy

TE – Transposable Elements

UPP – Ubiquitin Proteasomal Pathway

VMP – Volvox Matrix Metalloproteases

VMPs – Volvox Metalloproteases

List Of Symbols

μ - Mutation Rate

N_e – Effective Population Size

χ^2 – Chi-squared

R^2 – Correlation co-efficient of determination

K – K-mer

Glossary of terms

Genome: A complete set of genes or DNA of an organism or cellular location. Applies to nuclear DNA or organelle DNA.

Transcriptome: The complete set of RNA molecules for a given organism.

Proteome: The complete set of proteins for a given organism.

N50 – When ordered from largerest to smallest the N50 is the size of the largest contig or scaffold that when summed comprises 50% of the assembly. Therefore 50% of the assembly consists of contigs of scaffolds at least as large as the N50. The same applies other measures such as the N90 (10% of the assembly is larger) or N30 (30% of the assembly is larger).

Homologue: Gene related to another gene by common ancestry.

Orthologue: Gene from one species related to a gene in another species that originated from a speciation event.

Paralogue: Gene related to another gene that originated from gene duplication.

Fitness: The combination of the ability of an invidual to survive and reproduce. It is not a property but rather a probability for a population of individuals to survive and produce offspring.

Viability: The average ability for a population of individuals to survive.

Fecundity: The average ability of individuals from a particular population to reproduce and produce offspring.

Positive selection: In molecular biology positive selection is favouring of an advantageous genetic variant through the action of natural selection.

Neutral evolution: In molecular biology neutral evolution is the increase or decrease in frequency in a population of a genetic variant that offers no advantage or disadvantage (neutral variants) to organismal fitness.

Accelerating selection: The loss of purifying or negative selection, which may indicate positive selection or loss of function.

Purifying or negative selection: In molecular biology purifying selection is the removal of variants that are deleterious. Purifying selection results in sequence conservation. Sequences under purifying selection include constrained sequences, which if mutated result in the loss of function and have a deleterious effect on organismal fitness.

Evolvability: The capacity for an organism to evolve. In some definitions this may apply to the genetic diversity of a population such that the more variation the greater the probability is of a favourable mutation being present or arising. In definitions evolvability may instead refer to the innate properties of an organism to evolve into other forms.

Allometry: the relationship between organismal size and morphology/phenotype.

K-mer: A substring of the length k.

Shotgun assembly: Genome assembly by sequencing of small fragments of randomly sheared DNA with. Sequencing data is generated to be of sufficient volume such that the expected genome size was covered multiple times i.e. each base-pair in the genome may have been sequenced 10 times on average. Sequence data can then be aligned to produce a consensus sequence

Contig: Contiguous sequences generated from the alignment and assembly or overlapping sequence reads.

Scaffold: Multiple (>1) contigs linked together of known orientation with gaps of estimated length represented by ambiguous nucleotide bases ("N") between contigs.

Next generation sequencing (NGS): Sequencing methods performed massively in parallel generating gigabases or terabases of sequence data from single runs on an sequencing instrument. Data from NGS sequencers are typically short (100-300bp).

Paired-end sequencing: Sequencing of both ends of a DNA fragment producing linked pairs of sequence reads.

Mate-pair (MP) sequencing: Sequencing both ends of DNA fragments that are separated by large distance (e.g. 3-20kb). Typically, mate-pair sequences are generated by circularising large linear fragments of DNA and then exclusively capturing the region of the circularised molecule where the ends of the initially linear fragment joined. MP data are typically used to produce to sequence scaffolds.

Assembly algorithm: A computer program for producing a *de novo* genome or transcriptome sequence/s from sequence data.

Gene annotation: Producing a model of a gene structure that may include exon/intron boundaries, the determination of open reading frames as well as transcription and translation start and stop sites.

Gene functional annotation: Determining the function of a gene. Frequently annotating all genes in a genome this is performed through the identification of homology to genes with characterised functions.

Heterothallic: Where two genders are required to sexual reproduce. Also known as out-crossing.

Homothallic: The ability of a single organism to produce two genders that can sexually reproduce. Also known as selfing.

Gene ontology: A unifying hierarchical definition of gene function according to biological process (e.g. photosynthesis), molecular function (e.g. kinase activity) and cellular compartment (e.g. nucleus).

Domain: A conserved part of a protein sequence.

Gene Family: A collection of related genes. The relationship of genes may be defined by sequence similarity or by shared function. Sequence similarity is assumed to imply similar function. The most widely used meaning of a gene family is a collection of genes that are related by sequence similarity. There are no formal definitions for how genes are clustered together into families and various methods may be used to define a collection of genes as a family.

1 INTRODUCTION AND LITERATURE REVIEW

The evolution of multicellularity was essential for forming the aspect of life that we are most familiar with –macroscopic life. The volvocine green algae are amongst the best model lineages for exploring the evolution of multicellularity (Starr 1968, 1970; Kirk and Harper 1986; Kirk 1999, 2005; Herron and Michod 2008; Herron 2009). The volvocines span the gamut of cellular complexity and include related single-celled species (e.g. *Chlamydomonas reinhardtii*), colonial forms (e.g. the 4-celled *Tetrabaena socialis* and the 16-celled *Gonium pectorale*) and organisms consisting of hundreds to thousands of differentiated cells (e.g. *Volvox carteri*) (Nishii and Miller 2010; Coleman 2012). The multicellular volvocine algae evolved more recently than most other lineages of multicellular organisms (circa 209–260 million years ago) (Herron, et al. 2009) and there are many extant species of volvocines available for studying this evolutionary transition (Coleman 2012). Extant taxa with varying levels of morphological complexity are known and their genome sequences can be used to explore the molecular genetic underpinnings of the evolution of multicellularity and further increases in organismal complexity.

The volvocines are also a model lineage for understanding the evolution of sexual dimorphism because they range from isogamous to oogamous modes of reproduction (Martin and Goodenough 1975; Kirk 2006; Coleman 2012).

Volvocine gender is determined by multi-gene mating loci, which possess certain gender-limited or sex-determining genes (Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). However, both heterothallic and homothallic taxa are known in the lineage and the genetic differences between heterothallic and homothallic are not well understood.

In addition to the above areas of interest, the organelle genomes of volvocine algae are another subject of interest because of the large size of plastid genomes in taxa such as *Volvox carteri* and because of a complex history of transitions between linear and circular mitochondrial genomes in the lineage and related taxa (Michaelis, et al. 1990; Smith and Lee 2009; Hamaji, Smith, et al. 2013; Smith, et al. 2013; Del Vasto, et al. 2015). Volvocine organelle genomes have been used to explore nucleotide diversity and various analyses suggest that increases in organismal complexity are associated with decreased effective population sizes (Smith and Lee 2010; Smith, et al. 2013).

The nuclear genomes of *Chlamydomonas reinhardtii*, *Gonium pectorale* and *V. carteri* have previously been sequenced (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016) and this project adds the *T. socialis* genomes (nuclear and organelle genomes) to this body of knowledge.

In Chapter 2 findings from the organelle genomes of *T. socialis* are described. Both the mitochondrial and plastid genomes of *T. socialis* were found to be inflated with non-coding DNA. The plastid genome in particular at >400kb is amongst the largest of the known plastid genomes. The mitochondrial genome of *T. socialis* is circular, which together with data from other volvocine mitochondrial genomes, suggests that ancestral taxa that gave rise to the multicelled volvocines had circular mapping mitochondrial genomes. Transitions between linear and circular mapping mitochondrial genomes occurred many times in the volvocine lineage suggesting that the toolkit for processing either conformation is conserved. Furthermore, there is no evidence to date that mitochondrial genome conformation affected organismal fitness, which suggests that transitions have likely been driven by neutral evolutionary processes.

Chapter 3 forms the main body of work. It describes the techniques used to generate the *de novo* nuclear genome of *T. socialis* as well comparative genomics techniques employed to identify genetic developments associated with the evolution of one of the simplest forms of multicellularity. Additional results not included in Chapter 3 are included as appendices.

Findings from the nuclear genome of *T. socialis* include the identification of a small set of gene families that trace to the evolution of multicellularity as well as numerous genes with altered selection pressures

in the multicellular volvocines, in particular, genes related to the ubiquitin proteasomal pathway were associated with the evolution of multicellularity. The function and origin of gained gene families were explored and genes with functional assignments were found to be enriched in developmental functions while a significant subset were without homology to reference database proteins. Gained families were not enriched in proteins from other multicellular taxa. Together, genes gained at the origin of multicellularity suggest that the evolution of multicellularity in the volvocine lineage was due to lineage specific genetic developments. Altered selection pressures acting upon all steps of the ubiquitin proteasomal pathway (UPP) in the multicellular volvocines were identified, thereby implicating alterations to this pathway with the origin of multicellularity.

The UPP has a known role in many lineages for regulating the cell cycle and because a modified cell cycle is considered one of the earliest developments in the volvocine lineage it is most likely that altered selection pressures acting upon this pathway are associated with regulating the cell cycle.

In addition to the above findings, numerous gene families associated with developmental complexity in the volvocine lineage were examined in *T. socialis*. Analysis of *T. socialis* agrees with previous findings, which did not identify a correlation between the evolution of multicellularity in the volvocine lineage and increased protein-coding

complexity (Prochnik, et al. 2010; Hanschen, et al. 2016). Specific families related to the cell cycle, germ/soma differentiation, cell adhesion and the extracellular matrix were analysed and findings reported. Genomic and proteomic data for *T. socialis* is anticipated to be of value for the volvocine community.

Chapter 4 is a preliminary description of the *T. socialis* mating-locus. The mating locus sequence was not completed, however, two large scaffolds housing multiple mating-locus gene were identified. As the first homothallic volvocine taxon to be sequenced mating-locus genes provide the first insight into the molecular differences between heterothallic and homothallic volvocines. Genes such as *MID*, *FUS1* and *MTD1* are gender-limited in heterothallic taxa and all three of these genes were identified in the homothallic *T. socialis*. This chapter will be submitted for publication.

Premises: (i) Multicellularity in the volvocine algae evolved by the step-wise acquisition of distinct traits; (ii) most traits are reflected by genetic / genomic changes in key members in the lineage; (iii) specific genes/gene families are associated with specific morphological developments in the volvocine lineage.

Expected outcomes: (i); the organelle genomes of *Tetrabaena socialis* will provide insights into the volvocine organelle genome evolution with regards to organelle genome size evolution and mitochondrial genome

conformation; (ii) as a basal member of the volvocine lineage the nuclear genome of *Tetrabaena socialis* will aid in understanding the molecular underpinnings of the transition from unicellular living to colonial multicellularity; (iii) the mating-locus of the homothallic *Tetrabaena socialis* will serve to better understand the molecular genetic differences between homothallic and heterothallic reproduction in the volvocine lineage.

1.1 The evolution of complexity

The most striking feature of the history of life on earth is the seemingly inevitable progression of biological complexity at the macroevolutionary scale. The trend of increasing biological complexity over time is by no means a guaranteed outcome of Darwinian evolution and evolution can also lead to a decrease in complexity.

Biological systems seemingly resist decay into entropy and through processes of self-organisation act like the proverbial 'Maxwell's daemon'. Unlike complexity observed in inanimate systems, 'simple laws' do not govern biological systems – even if certain trends in biology have been labelled as laws/rules (e.g. Kleiber's law of allometric scaling (Kleiber 1932; Niklas 2006; Niklas and Kutschera 2015) and Hamilton's rule of altruism (Hamilton 1963)).

Biology is inherently stochastic in nature and dependent on historical contingencies (Koonin 2011; Ramsey and Pence 2016). To

‘replay’ the formation of the universe with the same governing laws of physics would most probably result in a universe with galaxies, stars, planets, and black holes. However, were life on earth to be ‘replayed’ it would likely look very different now. The underlying principles that resulted in the evolution of complexity remain contentious. Gould described the trend towards biological complexity as a ‘drunkard’s walk’ - an inevitable result of a moving system that is confined within boundary conditions (Gould 1996). Some argue that competition - ‘nature red in tooth and claw’ and ‘selfish genes’ (Dawkins 1976) – is sufficient to explain all of life’s complexity while others argue for more holistic models where cooperation and group fitness is central to the development of increasing complexity (Nowak 2006; Nowak, et al. 2010). Unfortunately, even defining and measuring biological complexity is challenging and contentious (Adami, et al. 2000; Adami 2002).

Colloquial understanding of complexity or the evolution of complexity often differs from that of an evolutionary biologist. The evolution of flight, for example, may be have been a complex evolutionary development in terms of the morphological innovation and neuromuscular coordination required for such an activity, but from the point of view of an evolutionary biologist an animal that flies is not inherently more complex than a non-flighted animal.

One way of defining biological complexity is through the application of ‘information theory’ (Adami, et al. 2000), which is a framework for computational problem solving. This is of some utility but fails to fully capture biological systems. Features such as genome complexity are a poor measurement of organismal complexity because although, when viewed across the broad phylogenetic spectrum of life, a trend towards greater genomic complexity correlates with other intuitive concepts of organismal complexity, there are simply far too many exceptions (Cavalier-Smith 2005). For example, the ‘humble’ *Amoeba* possesses a genome 30 times larger than the human genome (Cavalier-Smith 2005). Such discrepancies have given rise to the term known as the C-value paradox. Another measure of complexity is metabolic complexity (O'Malley and Powell 2016). In this regard prokaryotes and archaea display the greatest diversity and complexity, which seems to contradict our intuitive perception of biological complexity where multicellular eukaryotes are perceived to be more complex than prokaryotes.

Ultimately, a resolute definition of complexity that unites all aspects of biological systems may not be possible to formulate. The most widely used approach is to apply a hierarchical definition of life, stratifying life according to levels of cooperation and organisation that selection acts upon (Maynard Smith and Szathmáry 1995).

Eörs Szathmáry and John Maynard Smith identified 8 major evolutionary transitions (MTEs), which like steps on a metaphorical ladder make up the hierarchical levels of cooperation and organisation that selection acts upon. The MTEs were defined as "'major stages in the evolution of complexity that involve a change in the level of organisation, and hence the level of selection"' (Maynard Smith and Szathmáry 1995). This hierarchical definition of organismal complexity is not without criticism (O'Malley and Powell 2016) and, because complexity is not resolutely defined an intuitive appreciation of complexity should not be ridiculed. To paraphrase Potter Stewart: "'What is biological complexity? I'm not sure - but I'll know it when I see it"'.

Szathmáry recently redefined MTEs (labelled as MTE2) in an attempt to create an improved synthesis of MTE theory through placing greater emphasis on "'evolutionary transitions of individuality (ETIs)"' (Szathmáry 2015). Such ETIs may be defined as a transition in the target of selection from an individual to a higher-level individual (Michod and Roze 1997; Michod, et al. 2003; Michod and Herron 2006; Szathmáry 2015).

A prime example of this is the migration of the target of selection from single cells to a multicellular organism (Maynard Smith and Szathmáry 1995; Michod and Roze 2001). A greater understanding of ETIs is required at both the theoretical and mechanistic level. Interest in

how ETIs occur is not only a matter of academic interest, they relate to many aspects of practical importance. These include how conflict between lower-level units is mediated during transitions to higher levels of organisation as well the policing of 'cheating' behaviour by lower-level units once a higher level of organisation is established (Michod 1996; Michod and Nedelcu 2003). A practical example of the former aspect is the role of conflict mediation and cooperation for the emergence of human societies and social cohesion (West, et al. 2006; Scott-Phillips, et al. 2011). The latter aspect of policing is perhaps best illustrated by one of the most consequential examples of failed policing, which is cancer (Maley and Forrest 2000; Crespi and Summers 2005; Merlo, et al. 2006; Pepper, et al. 2009; Merlo, et al. 2010; Aktipis, et al. 2011; Heng, Stevens, et al. 2011; Featherston and Durand 2012; Aktipis, et al. 2015; Fortunato, et al. 2017). In cancer, selection reverts from the higher-level organism to the unicellular level through cellular competition for resources both amongst cancer cells and between cancer cells and 'healthy' cells (Heng, et al. 2010; Heng, Stevens, et al. 2011). Cancer may be described as the failure of multicellular cooperation.

1.2 The evolution of multicellularity

The first fossil evidence of multicellular eukaryotic life dates to over 2 billion years ago (El Albani, et al. 2010; Zhu, et al. 2016). Although the transition to multicellularity is a landmark development in organismal complexity, it has occurred independently at least 25 times (Wolpert and

Szathmary 2002; Niklas and Newman 2013; Niklas 2014), such as in metazoans, plants, green algae, red algae, brown algae, fungi, prokaryotes and amoebozoa (Knoll 2011; Herron, et al. 2013; Niklas and Newman 2013).

It is challenging to present a unifying definition of multicellularity because multicelled forms of life present in alternative configurations. Put simply, a multicellular organism possesses multiple cells linked together that are not readily separable (i.e. not aggregates or clumps), they typically pass through a single-cell phase in their life cycle, the cells within the organism share coordinated functions and the organism may possess division of labour (Niklas 2014; van Gestel and Tarnita 2017). However, many exceptions to these basic criteria are found and certain organisms meet some of these requirements but are missing others. For example, non-clonal communities of unicellular *Dictyostelium* cells can aggregate to form a transient multicellular organism with complex structures (Fortunato, et al. 2003; Smith, et al. 2014). If clonality and a single-celled bottleneck are a defining characteristic of a multicellular organism then *Dictyostelium* with its complex structures is not multicellular. Another challenge to the basic definition of multicellularity includes that certain multicellular organisms can revert to a single-celled existence and although they may mostly exist in a multicellular form they can be defined as facultatively multicellular as opposed to obligatorily multicellular. For such organisms selection has not been transferred fully from groups of cells to a higher-

level multicellular individual (Michod and Nedelcu 2003; Michod 2006). The simpler volvocine algae are an example of facultative multicellularity and are referred to as colonial rather than multicellular. Even the coordination of a collective function may be present in non-multicellular organisms, such as quorum sensing among bacteria (Diggle, et al. 2007). To circumvent formulating a strict definition of multicellularity and for fluidity of writing, for the remainder of this thesis the definition of a multicellular organism is an organism that possesses multiple cells of clonal origin.

It is acknowledged that this may not be a perfect definition and the inclusion of clonality excludes organisms such as *Dictyostelium*, however, the focus of this work is on the volvocine algae, which vegetatively reproduce clonally. The term colonial may be interchanged with multicellular throughout this work to denote simpler forms of facultatively multicellular volvocines in which selection has not fully been transferred to a higher-level unit. Undergoing a single-cell bottleneck producing clonal progeny is not only a trait for defining multicellularity but is a key factor that favours the emergence of multicellularity (Queller 2000). It is also the general case that genetic relatedness favours altruistic behaviour and cooperation (Hamilton 1963; Gardner, et al. 2011).

The division of labour through cellular specialisation is a hallmark of more complex forms of multicellularity (Willensdorfer 2009; Knoll 2011).

The most basic form of division of labour is where certain cells within a multicellular organism are capable of reproduction whereas other cells may relinquish this ability to exclusively perform somatic functions critical for survival and hence viability. This is known as germ-soma division of labour and it separates the two central aspects of fitness - viability and fecundity - into different cellular lineages. Depending on the definition, 'complex multicellularity' has evolved between 6 and 10 times in 5 major classes (Knoll 2011; Niklas and Newman 2013). Criteria for defining complex multicellularity include the possession of multiple cell types, tissues, organs, regulated forms of intercellular communication, a developmental programme and three-dimensional organisation (Knoll 2011). It may appear that the formation of complex multicellular organisms from simpler multicellular organisms was a greater barrier to overcome than the initial development of simpler multicellular organisms. However, complex multicellularity evolved ~6 times out of ~25 lineages, which is a high proportion when compared to the ~25 known cases of simple multicellular forms that evolved from a countless number of extinct as well as extant unicellular lineages (Umen 2014). The broader taxonomic perspective illustrates how challenging the initial transition to multicellularity is/was. The vast number of extant unicellular lineages is testament to the fact that evolution of increased organismal complexity will not always be favoured or guaranteed. Challenges faced during all ETIs include the mitigation of conflict amongst lower-level units and the control of 'cheaters' - lower-level units that exploit as opposed to contribute

towards the advantages that higher-level organisation may provide (Michod and Nedelcu 2003; Michod, et al. 2003; Santorelli, et al. 2008; Saxer, et al. 2010; Kuzdzal-Fick, et al. 2011; Strassmann and Queller 2011; Buttery, et al. 2013; Levin, et al. 2015; Inglis, et al. 2017). An excellent example of such 'cheaters' and the failure to police them is the example of cancer where low-level units (cells) reproduce in an uncontrolled fashion, exploiting resources (such as vascular systems and nutrient supply) and ultimately may lead to the demise of the higher-level collective (Crespi and Summers 2005; Heng, et al. 2010; Featherston and Durand 2012). Furthermore, increased organismal complexity and size correlates inversely with fecundity and metabolic rate (Lynch and Conery 2003; Niklas and Kutschera 2015). Reasons for this include reduced efficiency in nutrient/waste diffusion rates, intracellular transport rates, a reduction in the rate of motility (relative to size), decreased efficiency from scaling effects (allometry) as well as increases in factors such as requirements for nutrient accumulation prior to reproduction, which delays reproductive maturation (Niklas 2004; Niklas and Spatz 2004; Solari, Kessler, et al. 2006; Knoll 2011). A useful example of allometry, which is applicable to certain photosynthetic organisms, is the inverse correlation between organismal size and relative access to photons for photosynthesis (Niklas 2006; Beardall, et al. 2009). The possible negative consequences of increased complexity mandate that benefits in fitness driving increased complexity must exceed reduced rates of metabolism and fecundity.

It is also worth noting that whilst continued developments in complexity can potentially result in increased organismal efficiency through specialisation, natural selection is not prescient to potential future benefits and thus initial transitions to a higher level of organisation are likely to be subject to significant trade-offs between fecundity and viability (Michod and Nedelcu 2003; Solari, Kessler, et al. 2006; Leslie, et al. 2017).

Extant multicellular lineages are found across the phylogenetic spectrum. However, the ability to study the mechanistic and theoretical aspects of how multicellularity evolved in most lineages remains challenging. Such challenges include the vast time periods that have elapsed since transitions occurred (Sharpe, et al. 2015), difficulty in resolving phylogenies (Srivastava, et al. 2010; Kohn, et al. 2012; Moroz, et al. 2014; Halanych, et al. 2016), a lack of fossil evidence and the fact that species with intermediate properties are not known. The evolution of multicellular metazoans is an excellent example of these challenges. Today, even with genomes available for unicellular relatives of metazoans (King, et al. 2008; Suga, et al. 2013) as well as genomes from the multicellular porifera (Srivastava, et al. 2010), ctenophora (Moroz, et al. 2014), placozoa (Srivastava, et al. 2008), cnidaria (Putnam, et al. 2007) lineages and an extensive list of bilateria genomes (e.g. vertebrates) there remains controversy regarding which lineage is the most basal. Currently the ctenophora are considered the most basal of extant animal lineages

(Moroz, et al. 2014; Whelan, et al. 2015). The vast time period (~600-1,200 MYA) that has elapsed since animal multicellularity evolved means that extant colonial (simple) animals are probably extinct or unknown. Furthermore, a reconstruction physiological and anatomical developments as well as biochemical and molecular pathways present in the first multicellular animal has proven challenging and the most parsimonious account of developments or losses is puzzling (Moroz, et al. 2014). For example, the ctenophora are considered basal to other animal lineages, including the porifera and placozoa, however, neurones, an elementary brain and muscles are present in the ctenophora but not in the porifera and placozoa. In the chlorophyte volvocine lineage many of these challenges are less constraining and this lineage offers an excellent model for studying the evolutionary transition to multicellularity.

1.3 The Volvocine Lineage

The volvocine green algae are an excellent model lineage for exploring the origins and evolution of multicellularity. At the level of whole genome data, three data sets are available (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016) (the fourth whole genome project was the main component of this thesis). Its members span the gamut of cellular complexity and include the single-celled species sister to the multicelled taxa (e.g., *Chlamydomonas reinhardtii*), colonial forms (e.g., the 4-celled *Tetrabaena socialis* and the 16-celled *Gonium pectorale*), organisms with partially differentiated cell types (e.g.,

Pleodorina starrii), and complex multicellular taxa with germ-soma differentiation (e.g., *Volvox carteri*) (Nishii and Miller 2010; Coleman 2012). In terms of morphology species within the lineage range from unicellular to flat sheets of cells to spherical forms. It is estimated that colonial volvocine algae arose between 209–260 million years ago (MYA), after which the lineage underwent rapid radiation, giving rise to numerous multicelled taxa (Herron, et al. 2009).

The multicellular volvocine algae form a monophyletic lineage (Larson, et al. 1992; Coleman 1999; Nozaki, et al. 2000; Nakada, et al. 2008; Herron, et al. 2009) of over 50 species within the Reinhardtinia clade of the Chlamydomonadales order (class Chlorophyceae) (Nakada, et al. 2008) (Figure 1.1). In addition to the multicellular volvocines, the term volvocine may include the unicellular *C. reinhardtii* and other related unicellular genera such as *Vitreochlamys*, *Polytomella* or other members of the reinhardtinia-clade (Nakada, et al. 2008). The phylogenetic relationships among volvocines have been investigated extensively (Larson, et al. 1992; Coleman 1999; Nozaki, et al. 2000; Nakada, et al. 2008; Nozaki, et al. 2014), and the multicellular taxa are currently divided into three families: the Volvocaceae (e.g. *P. starrii* and *V. carteri*), the Goniaceae (e.g. *G. pectorale*), and the Tetrabaenaceae (e.g. *T. socialis*), which is the deepest-branching of the three families (Nozaki, et al. 2000). It is noted that phylogenetic evidence suggests that a number of clades may exist amongst the Volvocaceae, which may prompt further

subdivision of this family (Coleman 1999, 2012; Nozaki, et al. 2014). Both the *Chlamydomonas* and *Volvox* genera are polyphyletic because their initial descriptions were based for the most part on morphology. Species of *Volvox* such as *V. carteri*, *V. obversus*, *V. africanus*, *V. gigas* and others are frequently portrayed as the most developmentally complex in the lineage (Viamontes, et al. 1979; Kirk 1988; Kirk 1998; Kirk and Nishii 2001; Miller 2002; Kirk 2003, 2005; Herron and Michod 2008; Nishii and Miller 2010; Matt and Umen 2016). This is because the various species of *Volvox* are obligatorily multicellular and are therefore fully committed to group living, cell number in some species can exceed 50 000, all species possess germ-soma division of labour, and *Volvox* reproduces sexually via oogamy (Cave and Pocock 1951; Coleman 2012).

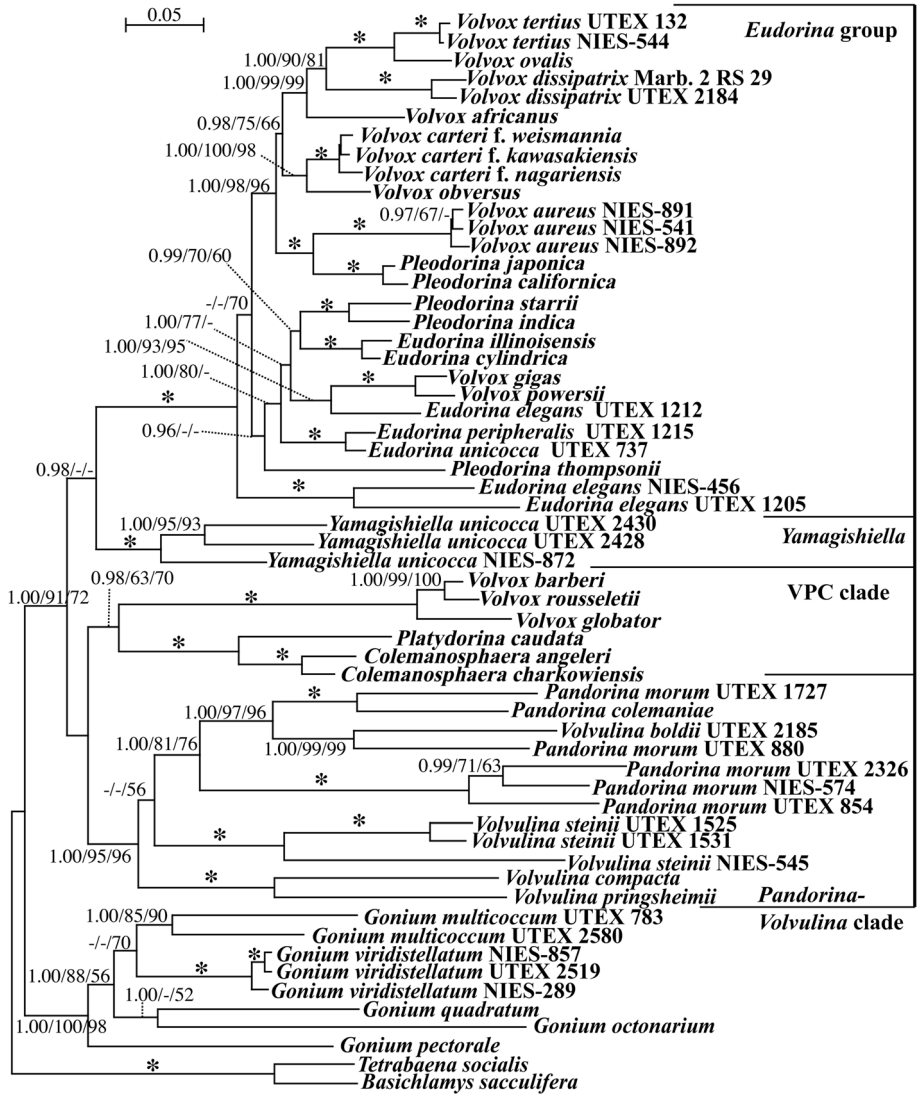


Figure 1.1: A recent phylogenetic construction of volvocine relationships. The three families of volvocines, the Tetrabaenaceae, Goniaceae and the Volvocaceae, form distinct branches. Further subdivisions within the Volvocaceae as well as the polyphyletic state of the *Volvox* genus are shown. Image taken with permission from (Nozaki, et al. 2014).

Molecular sequence analysis has demonstrated a close relationship of the unicellular *C. reinhardtii* to the multicellular taxa and *C. reinhardtii* is frequently used as a proxy for the ancestral unicellular alga that gave rise to the multicellular forms. The close relationship of *C. reinhardtii* to the multicellular taxa is also particularly advantageous for studying the evolution of multicellularity because *C. reinhardtii* is probably the most well studied chlorophyte and is a model species for studying a range of traits such as cellular motility, photosynthesis, metabolism, the cell cycle, lipid production etc. (Goodenough and Levine 1974; Steinback and Goodenough 1975; Martin, et al. 1976; Goodenough and Forest 1978; Janero and Barrnett 1981; Michaels and Herrin 1990; Merchant, et al. 2007; Dillard, et al. 2011; Blaby, et al. 2013; Didur, et al. 2013; Blaby, et al. 2014; Goodenough 2015; Dinc, et al. 2016).

While the volvocine lineage is a prime model for studying the evolution of multicellularity, there are general weaknesses in studying extant taxa to understand how multicellularity evolved historically. The most fundamental error made is that extant taxa are inferred to be the same as an ancestral organism. This is particularly true if no fossil records are available, such as for the volvocines. With over 200 million years of separation it is impossible that volvocine taxa that possess intermediate properties or traits have not themselves undergone modification in ways that we cannot necessarily know.

Extant members of the volvocine lineage have been lined-up and portrayed as a simple series of progressive developments towards complexity (Kirk 2005; Nishii and Miller 2010; Hallmann 2011). However, such progressive depictions of evolution are an oversimplification of how nature has proceeded (authors may acknowledge that this is an oversimplification). For example, some now consider that the last common ancestor of humans and chimpanzees did not possess extended forelimbs such as those of modern chimpanzees (White, et al. 2014). The idea that chimpanzees have remained evolutionarily stagnant is now dismissed. Similarly, many species of extant volvocines with intermediate properties may in fact have lost developmental traits. Phylogenetic reconstruction points towards loss of specialised germ cells in *Volvox dissipatrix* and loss of other developmental features in species of *Pandorina* and *Eudorina* (Herron and Michod 2008; Herron, et al. 2009). Again, this paints a more complex picture than a linear progression of developments. Extant 'intermediate species' should more correctly be considered as species with intermediate traits that may be well suited to their environment and such organisms may have undergone significant modifications since divergence from other taxa. This does not, however, nullify attempts to better understand the causes of trait development.

Some have questioned the relevance of studying historical developments of complex traits, such as multicellularity, through the study of extant lineages and prefer instead to study complex traits using

experimental evolution (Murray 2013). However, there are also weaknesses with experimental evolution approaches, not the least being that they do not reflect what has occurred in nature (artificial selection provides a different and misleading account of natural selection). The stochastic nature of evolutionary processes makes an exact experimental replication of the path taken by nature impossible. This limits experimental evolution to only providing insights into general processes, trends or principles without revealing what may have occurred naturally.

1.3.1 Brief introduction to Volvocine life cycles

Members of the volvocine lineage can undergo sexual and asexual life cycles (Nishii and Miller 2010). The trigger for entry into the sexual life cycle differs between *C. reinhardtii* and *V. carteri*, while the trigger for the sexual life cycle is not well described in other multicellular taxa (Stein 1966; Nozaki 1986). In *C. reinhardtii* the sexual life cycle is triggered by N-deprivation (Harris 1989) while a pheromone induces the *V. carteri* sexual life cycle (Tschochner, et al. 1987; Mages, et al. 1988; Hallmann, et al. 1998). The sexual life cycle is described in Chapter 4.

Most of the colonial volvocine algae reproduce via palintomic multiple fission division, where a large mother cell grows until division is triggered and subsequently multiple rounds of mitotic divisions ensue in rapid succession where growth does not occur between divisions (Umen and Goodenough 2001; Cross and Umen 2015). This form of division is

less common than binary fission (e.g. budding), however, it is a common form of division among unicellular eukaryotes (Cross and Umen 2015). Synchronised cultures of *C. reinhardtii* follow a 24-hour life cycle where the growth phase occurs during daylight and division/mitosis occurs at night (Howell, et al. 1977; Wilson and Chiang 1977; Spudich and Sager 1980). The number of mitotic divisions is not genetically programmed but rather is regulated by the size of the mother cell entering into the division phase (Cross and Umen 2015; Li, et al. 2016; Shelton, et al. 2017). During division, flagella are lost and the dividing cells are non-motile. For a brief time period cells that have divided remain connected within the outer cell wall of the mother cell and this phase might be described as a transient multicellular phase (Maliet, et al. 2015). After this brief phase, daughter cells cleave, hatch from the mother cell and regrow flagella.

The asexual life cycle of *V. carteri* proceeds over approximately 48 hours (Kirk 2001). *V. carteri* cell division is restricted to differentiated germ cells (n=~16), which are defined during embryogenesis (Kirk, et al. 1991; Larson, et al. 1992; Kirk 2001). Somatic cells are incapable of further division while during the growth phase germ cells achieve a large cell volume capable of dividing 12-14 times (Kirk 1988, 2005; Hallmann 2011). During the division phase of the life cycle germ cells divide symmetrically (equally sized daughter cells) until a 32-celled embryo is formed (Schmitt 2003), after which, the division programme transitions to an asymmetrical programme of division in one half of the embryo, which results in fewer

large cells and more of the smaller cells. At this point a cell's fate has been determined. Large cells will go on to become germ cells and smaller cells become soma. In certain species of *Volvox* divisions may occur over a longer period of time and cells may grow in between divisions (binary fission) (Solari, et al. 2008), however, *V. carteri* divisions always occur via palintomy. At the end of the division phase the developing embryo is arranged with motile soma facing inwards and germ cells located on the exterior of the embryo. In a process akin to metazoan gastrulation the embryo inverts in order for flagella to point outwards and for non-motile germ cells to move to the interior (Viamontes and Kirk 1977; Nishii, et al. 2003; Nishii and Miller 2010). In *V. carteri*, after inversion cytoplasmic bridges that held cells together during inversion are replaced by an extracellular matrix of glycoproteins, which thereafter replaces the role of inter-cellular connectivity as well as making up the bulk of organismal volume. A period of growth within the mother colony ensues until the final stage of embryo maturation where gonidia and soma mature and, after which, no further growth takes place within somatic cells (Kochert 1968). After embryo maturation the embryos hatch from the outer organismal sphere, parental somatic cells die and the now fully formed hatchlings begin to grow again by the deposition of ECM (Kochert 1968).

The life cycles of *Chlamydomonas* and *Volvox* might be contrasted as the 'ancestral' unicellular programme versus that of one of the most derived programmes in the volvocine lineage, respectively (Figure 1.2).

Between these two programmes there are a number of alternative division programmes in other volvocines, some of which might be considered as 'intermediate'. Similar to *C. reinhardtii*, all cells from volvocines that lack germ-soma differentiation begin as motile and then later lose motility and enter a reproductive mode. The division programme of *Pleodorina starrii* is of particular interest, as it possesses partial germ-soma differentiation (Kirk 2005). In *P. starrii*, all cells begin as motile but later in the life cycle only some cells become reproductive.

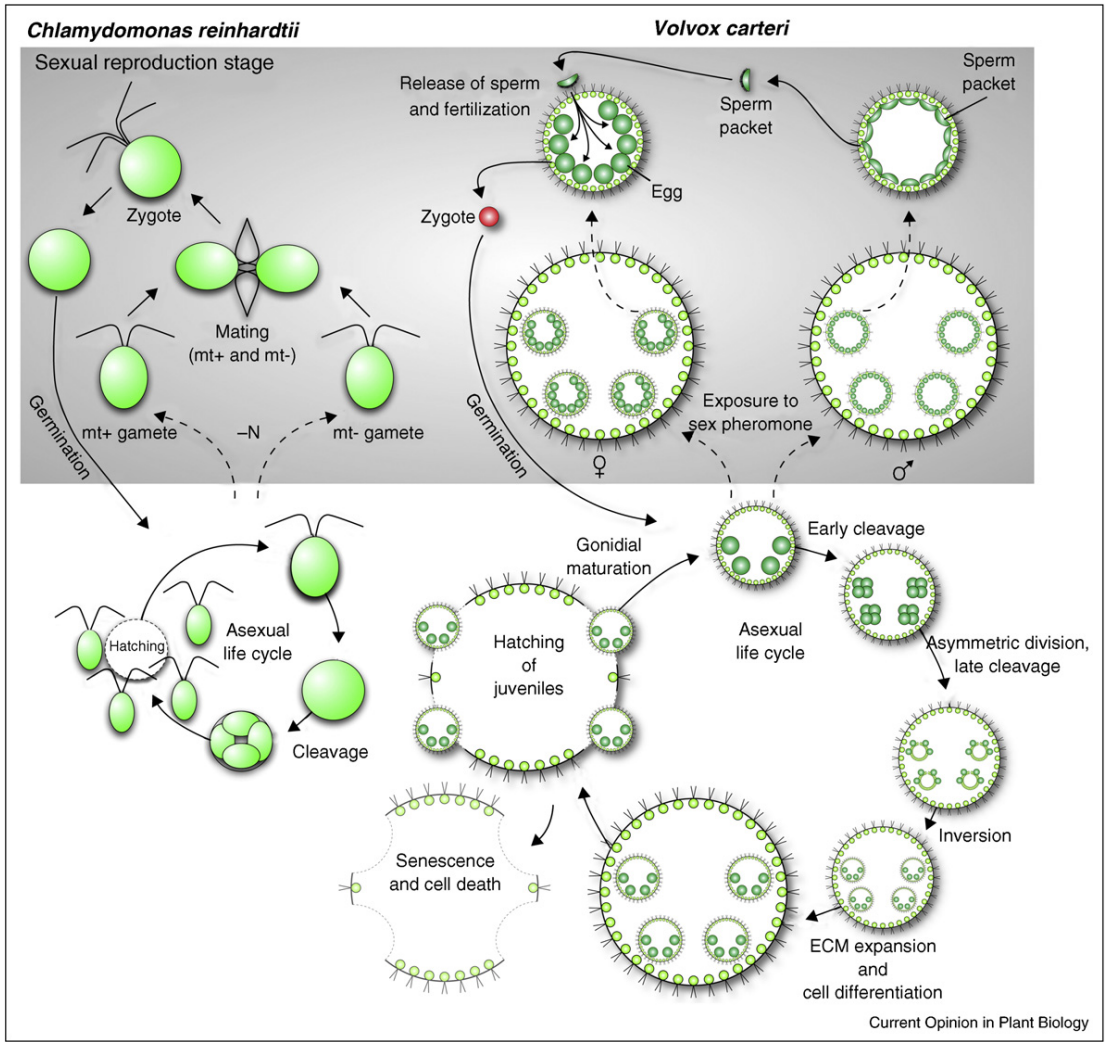


Figure 1.2: The respective life cycles of the unicellular *Chlamydomonas reinhardtii* and *Volvox carteri*.

Life cycles are described in detail in the text. Important points to note are: (i) that both organisms can reproduce either sexually or asexually, (ii) the size equality of *Chlamydomonas* gametes versus the different sizes of *Volvox* motile sperm and immotile eggs, (iii) the development of *Volvox* gametes within the mother organism, (iv) the death of *Volvox* soma after embryo hatching, (v) the loss flagella during *Chlamydomonas* cleavage and (vi) the point during the *Chlamydomonas* life cycle where cells are

cleaved but remain connected prior to hatching. Image taken with permission from (Nishii and Miller 2010).

1.3.2 Kirk's 12 steps to multicellularity

Understanding the molecular basis for multicellularity in the volvocine lineage was greatly advanced by David Kirk and co-workers (Viamontes and Kirk 1977; Viamontes, et al. 1979; Green and Kirk 1981; Green, et al. 1981; Green and Kirk 1982; Kirk, et al. 1986; Kirk and Harper 1986; Harper, et al. 1987; Kirk 1988; Tam and Kirk 1991b; Tam, et al. 1991; Larson, et al. 1992; Kirk, et al. 1993; Kirk 1994, 1997; Kirk, et al. 1999; Hallmann and Kirk 2000; Kirk and Nishii 2001; Kirk 2003; Nishii, et al. 2003; Kirk 2005). Amongst these contributions was the conceptual framework developed for understanding the evolution from a unicellular alga to the *V. carteri* form (Kirk 2005). The framework was derived from studying the phylogenetic, morphological and developmental characteristics of extant volvocines and the developmental route to the *Volvox* form was divided into 12 key steps.

Kirk acknowledged that the simple progression seen in extant species was an oversimplification, however, the framework is important because it provided a reasonable explanation for how the transition to a *Volvox carteri* form might have occurred through a step-wise series of smaller developments (Kirk 2005). The framework was later revisited and

refined by Herron and Michod (2009) using updated phylogenetic data for the reconstruction of ancestral character states (Figure 1.3). The framework may require further revision subsequent to a detailed morphological description of *Tetrabaena socialis* (Arakaki, et al. 2013).

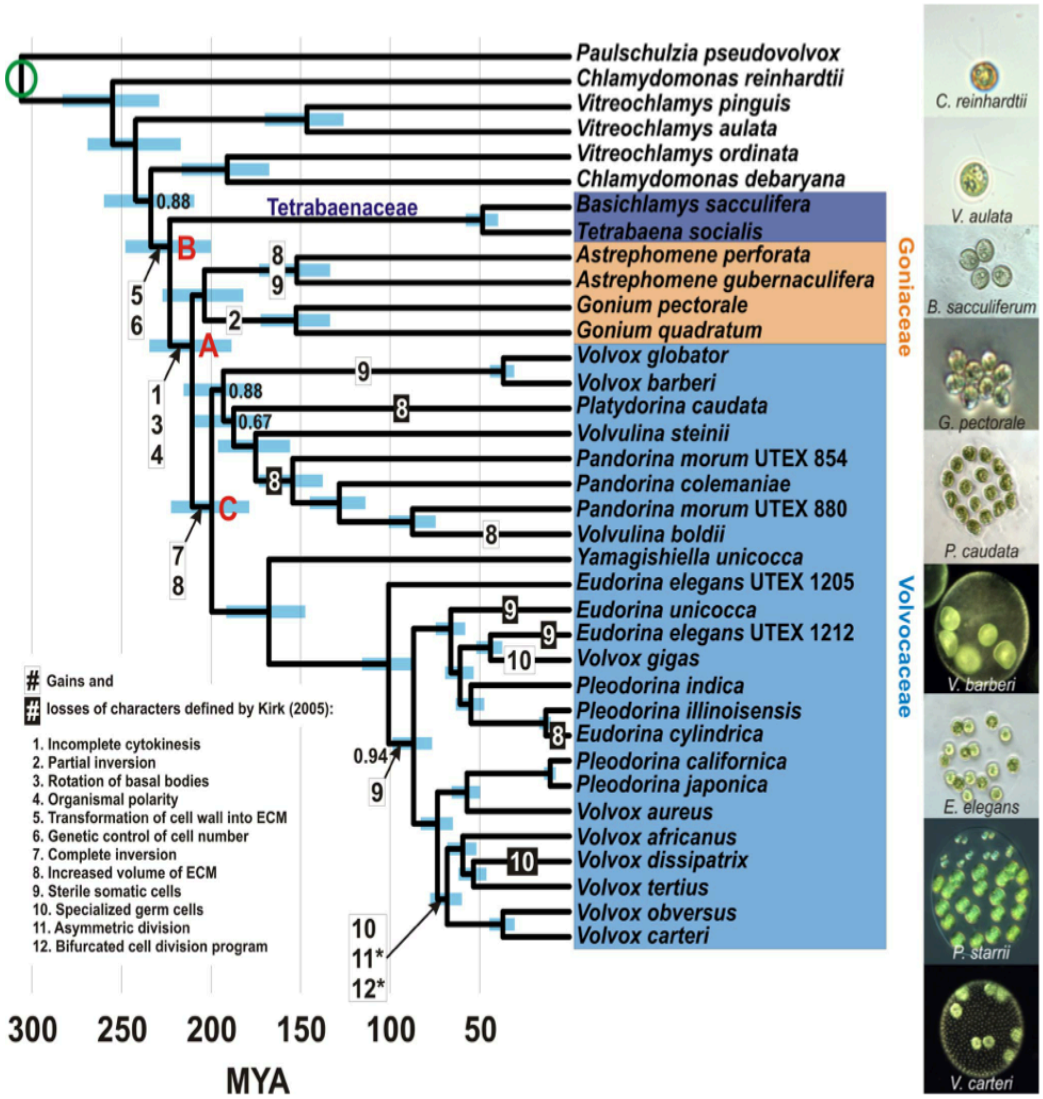


Figure 1.3: Ancestral character states of Kirk's 12 steps plotted against a volvocine phylogeny. On the right of the figure are images of selected volvocine chlorophytes. The timeline presented in the figure is measured in millions of years (MYA) and was derived using molecular

dating with fossil calibration. Examples of character losses as well as independent character gains are presented. For example, sterile soma (trait 9) may have been lost twice and independently gained 3 times. More recent evidence suggests that at steps 1 and 3 originated at node B in the phylogeny (Arakaki, et al. 2013). Image taken with permission from (Herron, et al. 2009).

Kirk's 12 steps are as follows:

1.3.2.1 Incomplete cytokinesis

During multiple fission division in *C. reinhardtii* complete cytokinesis separates progeny into individuals while in colonial and multicellular volvocines this final separation is incomplete and daughter cells remain connected by cytoplasmic bridges (in some species of *Volvox* these cytoplasmic bridges are lost in the mature organism). Incomplete cytokinesis may occur occasionally in *Chlamydomonas spp* but only infrequently (Coleman 2012).

1.3.2.2 Incomplete Inversion of the Embryo

With respect to the orientation of gonidia and somatic cells immature *Volvox* embryos are in a reverse orientation such that gonidia are positioned on the outside of the immature embryo and somatic cell flagella are facing inwards. In a process similar to metazoan gastrulation the developing embryo inverts around cytoplasmic bridges into the correct

orientation. In some species of *Gonium* a partial form of inversion is evident. Early *Gonium* embryos form a flat concave bowl-shaped sheet with flagella ends positioned within the hollow of the bowl shape. A partial inversion process later drives the bowl inwards to create a convex shape with the flagella ends facing outwards (Iida, et al. 2013). Intriguingly, in *Astrephomene gubernaculifera* (a member of the Goniaceae) a spherical body plan in the mature colony forms without the need for inversion (Yamashita, et al. 2016).

1.3.2.3 Rotation of Basal Bodies

The two basal body structures that anchor flagella in *C. reinhardtii* are orientated 180^{o} from one another to create a 60^{o} angle between the flagella. This allows for 'breast-stroke' flagella motion. The same is true for the central cells in *G. pectorale*. In peripheral *G. pectorale* cells and in other multicellular volvocines basal bodies are aligned to create parallel flagella (Iida, et al. 2013).

1.3.2.4 Establishment of Organismal Polarity

In *G. pectorale* the difference in basal body orientation and cell shape between central and peripheral cells is a simple form of organismal polarity (Figure 1.4)(Iida, et al. 2013). In all spherical body-plan volvocines that undergo inversion (and in *Astrephomene gubernaculifera*) an anterior to posterior body plan is in place. For example, the 4 sterile somatic cells are located in the posterior end of *A. gubernaculifera* and in most

Volvocaceae the larger germ cells are concentrated in the posterior (Figure 1.4) (Yamashita, et al. 2016).

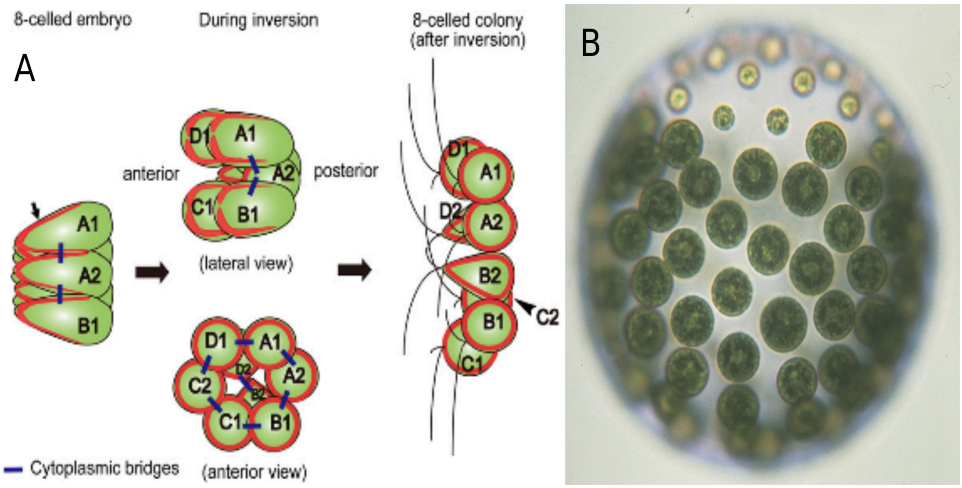


Figure 1.4: Central to Peripheral as well as anterior to posterior organismal polarity

(A) Lateral and anterior views of *Gonium pectorale* at different stages of embryogenesis. Anterior cells are rounder in shape and flagella are in parallel configuration. Central cells are elliptical in shape and flagella are position in anti-parallel configuration. Image taken from Iida, Ota and Inouye (2013). (B) *Pleodorina japonica* showing the accumulation of gonidial cells in the middle to posterior of the organism. Image B kindly provided by Prof Hisayoshi Nozaki.

1.3.2.5 Transformation of Cell Walls into an Extracellular Matrix (ECM)

C. reinhardtii possesses a 2-layered cell wall (inner and tripartite outer layer). A similar outer cell layer does not surround individual cells in *Volvox* and other larger volvocines. In *V. carteri* an extracellular matrix

(ECM) composed of glycoproteins instead holds cells together and the outer layer surrounds the organism instead of individual cells. In the smaller colonial taxa, such as *T. socialis* and *G. pectorale*, a tripartite layer surrounds each cell together in addition to an ECM (Kirk 2005).

1.3.2.6 Genetic Modulation of Cell Number

In *C. reinhardtii* division occurs through palintomic multiple fission division and the number of progeny produced is variable, which is determined by the size of the mother cell – typically ranging from 1 to 5 divisions (Kirk 2005). In contrast to *C. reinhardtii*, multicellular volvocines generally possess a characteristic maximum number of cells and maximum number of divisions. The maximum number of divisions in *T. socialis* is 2 whereas in *Volvox spp* the maximum number of divisions may reach as many as 14-16 divisions. Control of division number is genetically programmed in multicellular volvocines, a feature of all extant multicellular volvocines and hence the trait is considered to have emerged early in the lineage (Herron and Michod 2008; Herron, et al. 2009). However, the actual number of divisions is organism specific, which possibly indicates that the number of divisions is controlled by quantitative regulation.

1.3.2.7 Complete Inversion of the Embryo

As already discussed under heading 1.3.2.2, in the majority of spherical forms (a notable exception being *A. gubernaculifera*) the developing embryo undergoes complete inversion to orientate flagella

ends toward the exterior of the organism and to place germ cells in the interior. The process can be illustrated by imagining a tennis ball that is 'inside-out'. To invert the ball a cut is made in the ball and the ball is forced into the normal configuration with fluff (akin to flagella) on the outside of the ball (Figure 1.5).

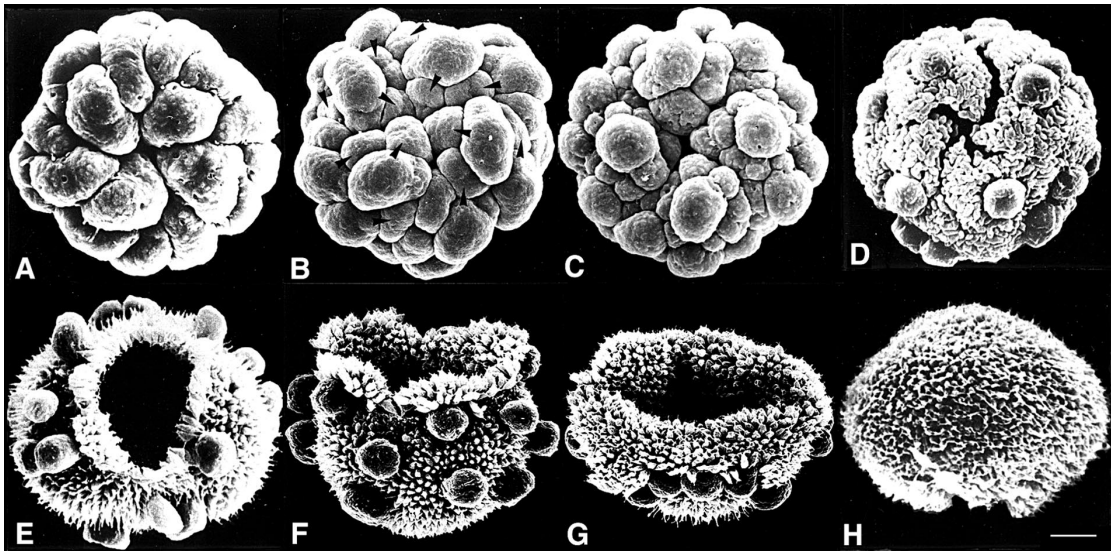


Figure 1.5: Electron micrograph images of the developing *Volvox* embryo undergoing inversion. In the images A through to E the position of gonidial cells on the exterior of the immature embryo can be clearly seen. In addition, flagella are facing inwards. In images D and E a slice through the embryo develops and expands. Images F and G show the process of embryo inversion taking place. In image H the embryo has inverted, gonidia are located in the interior and flagella point outwards. Image taken from (Kirk and Nishii 2001).

1.3.2.8 Increased Volume of the ECM

The ECM makes up a substantial portion of colonial volvocines, however, in *Volvox spp* the ECM is expanded and can comprise ~99% of the organism by volume. Throughout the lineage the ratio of ECM to cellular volume correlates with cell number. The ECM takes on the role in *V. carteri* of imbedding and connecting cells in a gelatinous mesh that is also presumed to serve as a nutrient depository. The ECM in *V. carteri* is divided into 3 zones known as the flagella zone, boundary zone and deep zone (Coleman 2012).

1.3.2.9 Partial Division of Labour

In colonial genera such as *Tetrabaena*, *Gonium* and *Pandorina* all cells have the same differentiation potentional exists and all cells are capable of undergoing division. For most of the life cycle cells are motile soma and then prior to division they lose flagella and convert to a reproductive mode. In *Pleodorina starrii* all cells begin undifferentiated and then later only a subset of cells differentiate into reproductive cells (Kirk 2005; Herron and Michod 2008). This life history is considered as an intermediate or partial form of division of labour because while division of labour is present the germ/soma differentiation occurs in the adult organisms and is not fixed during embryogenesis.

1.3.2.10 Complete Germ and Soma Division of Labour

In *V. carteri*, division of labour and cell differentiation into germ and soma is determined during embryogenesis and hence this differentiation is already present as the fledgling *Volvox* hatches from the mother. This is described as complete germ/soma division of labour.

1.3.2.11 Asymmetric Division

Cell size is the major factor that defines which cells in developing *V. carteri* (and for example *V. obversus*) embryos will become either motile soma or reproductive germ cells. Divisions begin symmetrically but after 5 symmetrical divisions further divisions are asymmetrical. The division plane no longer bisects the goniadial cells into equal sized progeny. The larger cells (>8μm) go on to become gonidia whereas the smaller cells become sterile somatic cells.

1.3.2.12 Bifurcation of the Cell Division Programme

Once defined as either germ or soma through asymmetrical division germ cells divide a further 3 times while somatic cells divide 6-7 times. The alternative division programme for germ and soma is known as the bifurcation of the cell division programme.

Kirk's framework was re-evaluated by Herron and Michod (2009) through reconstructing ancestral character states against a more extensive phylogeny. Furthermore, a phylogeny of photosynthetic

eukaryotes (with fossil calibrated divergence times) was used by Herron et al. (2009) to associate molecular dates to key developments in the volvocine lineage. This analysis dated the initial transition to ~234 MYA (209-260MYA), which is more ancient than previous estimates of 50MYA.

The split between arguably the most derived of volvocines - *Volvox carteri* (all 12 steps) - and near relatives *Volvox tertius* /*africanus* (10 steps) occurred possibly as recently as 35 MYA (Herron, et al. 2009). The more inclusive and formalised phylogenetic reconstruction of ancestral states by Herron and Michod (2009) is largely in agreement with the model proposed by Kirk, however, the step-wise developments were found to be more nuanced than initially proposed. An intriguing outcome of Herron and Michod's work was that some of the more derived volvocine characteristics were identified as independently gained or lost at various points in the phylogeny.

Independent evolution of certain traits (also known as convergent evolution) in the lineage includes the expansion of the extracellular matrix, the evolution of soma and the evolution of germ cells (Herron and Michod 2008). The evolution of soma may have arisen as many as three times in the lineage. Kirk also described convergent traits. Kirk identified that the first 5 steps of the transition were present in all extant volvocines. Through examination of extant organisms Kirk concluded that organisms without all of the first 5 steps present were likely to have impaired motility and

consequently were unlikely to survive (Kirk 2005). He concluded that colonial organism that gave rise to the volvocine lineage probably evolved all five steps simultaneously and that this leap explains the rarity of multicellular organisms amongst chlorophytes. Herron and Michod (2008) disagreed with this conclusion. Of the first 5 steps they found that the steps 'incomplete cytokinesis', 'rotation of basal bodies' and 'organismal polarity' were absent from the Tetrabaenaceae and concluded that these were later developments in the Goniaceae and Volvocaceae (the traits 'transformation of the cell wall into an ECM' and 'control of cell number' were considered to be present in the Tetrabaenaceae). This disparity is partly due to the fact that Kirk did not consider the Tetrabaenaceae in his analysis. However, a more recent morphological description of *T. socialis*, which until recently was not described in great detail, supports Kirk's initial conclusion.

With a thorough ultrastructural examination Arakaki et al (2013) found *T. socialis* to possess the first 5 of Kirk's 12 steps ('incomplete cytokinesis', 'rotation of basal bodies', 'establishment of organismal polarity', 'transformation of cell walls into an ECM' and 'genetic modulation of cell number'). It must be noted, however, that organismal polarity in *T. socialis* is distinct to that of *G. pectorale* (central-peripheral polarity) and, for example, *V. carteri* (anterior-posterior polarity). Therefore, Kirk's assertion that a multicellular volvocine requires all 5 steps might have been refuted because organismal polarity as defined by Kirk most likely

arose after the divergence of *T. socialis* from the ancestors of taxa such a *G. pectorale* and *V. carteri*. Recent studies of experimental evolution in *C. reinhardtii* that have attempted to re-create the evolution of multicellularity (Ratcliff, et al. 2013) have resulted in immotile colonies (Boyd, et al. 2018). Applying selective pressure for increased size produced multicellular colonies, however, the resulting colonies were non-motile. In this regard the experimentally derived multicellular colonies do not resemble known volvocines. In the volvocines motility might be considered as the major constraint faced during the transition to multicellularity (Solari, Ganguly, et al. 2006; Solari, Kessler, et al. 2006).

1.3.3 Driving forces behind Increasing Complexity in the Volvocine Algae

As already discussed, theory as well as observations from across the phylogenetic spectrum confirm that the transition to simple multicellularity as opposed to the development of complex multicellularity was a greater hurdle to overcome. Unicellular life remains by far the most common life-history strategy, which suggests that for most lineages and in the majority of ecological niches the evolution of multicellularity has either not been selected for and that there are challenges in evolving multicellularity. This raises the question as to *why* multicellularity has repeatedly evolved. The first or primary transition to multicellularity is not the subject of this thesis but rather an investigation into the evolution of multicellularity in the context of an already 'multicellular world'. Again, the

volvocines are an excellent model for studying evolutionary constraints during secondary transitions to multicellularity. As an historical event the question of *why* multicellularity and further complexification evolved in the volvocines is unresolved (Solari, Ganguly, et al. 2006; Solari, Kessler, et al. 2006).

Often-cited reasons for the evolution of secondary multicellularity that are relevant to the volvocine lineage include; protection from predation, greater rate of motility, nutrient storage as well as improved nutrient diffusion (Bonner 1998; Boraas, et al. 1998; Solari, Kessler, et al. 2006; Sathe and Durand 2016; van Gestel and Tarnita 2017).

The most well studied and compelling cause for increased organismal size in chlorophytes is predation (Boraas, et al. 1998; Solari, Kessler, et al. 2006; Fisher, et al. 2016; Kapsetaki, et al. 2016; Sathe and Durand 2016) and, intriguingly, various alternative strategies were employed by chloropyta to increase size (Umen 2014). Algae are vulnerable to predation from simple predatory flagellates such as euglenoids (Sathe and Durand 2016) and therefore the pressure from predation may have been pivotal for the development of multicellularity in the volvocine lineage (Solari, Kessler, et al. 2006). In experimental settings predation has been used to induce multicellularity or aggregation in chlorophytes (Boraas, et al. 1998; Kapsetaki, et al. 2016; Sathe and Durand 2016). However, certain chlorophytes have achieved tremendous

body size simply via increasing cell size. Extreme examples of which include, *Acetabularia* (single nucleus chlorophytes), coenocytic algae and siphonocladous algae (Niklas, et al. 2013; Umen 2014). What is curious and remains unexplained is that once a certain threshold size was evolved that is protective from predation, that size continued to increase in the volvocines. *Volvox carteri*, for example, is significantly larger than what small predatory flagellates can typically engulf. The same is true for many of the smaller spherical volvocines. Other factors that might explain this continued increase in organismal growth include increased storage of limiting nutrients like phosphate and an increased rate of motility, the latter of which is beneficial for both phototaxis and chemotaxis (Solari, Ganguly, et al. 2006).

Motility is critical for the volvocines and retaining motility during increases in complexity was obligatory. Solari et al (2006) highlighted the key role of motility during the evolution of volvocine complexity and furthermore, successfully modelled motility as a proxy for viability. Solari, Kessler et al. (2006) modelled life-history trade-offs in the volvocines between the two most basic of fitness functions - viability and fecundity. Through experimentation, mathematical modelling and by applying the principles of hydrodynamics motility was identified as the key driver for the evolution of germ and soma (Solari, Kessler, et al. 2006). The models derived by Solari and co-workers predict that increases in cell number can be detrimental towards growth and viability (motility) in volvocines without

cell specialisation and that at a certain point cell specialisation becomes a necessity. Dedicating certain cells to division mitigates the cost of increased cell number and its effect on motility. The models correlate remarkably well with the observed morphologies of extant volvocine taxa, including the absolute limit of cell number for the spherical volvocine form and the ratio of germ to soma for optimal motility at varying levels of cell number. This model is highly suggestive that the development of volvocine germ-soma differentiation emerged for adaptive reasons.

1.3.4 Genetic Developments within the Volvocine Lineage

The genetic underpinnings of several of Kirk's 12 steps have been identified using biochemical and chemical assays, targeted molecular approaches, proteomic techniques, whole-transcriptome experiments, knockout experiments and whole-genome sequencing. To date, the genomes of *C. reinhardtii* (Merchant, et al. 2007), *G. pectorale* (Hanschen, et al. 2016) and *V. carteri forma nagariensis* (Prochnik, et al. 2010) have been made publically available and additional genomes are currently being sequenced. Key findings as well as knowledge gaps from the aforementioned studies are discussed below and are aligned with certain developments in the lineage.

1.3.4.1 Incomplete Cytokinesis

The molecular mechanisms that underlie incomplete cytokinesis are as yet unknown, however an important morphological feature of

multicellular volvocines is the cytoplasmic bridges that maintain connections between cells. Cytoplasmic bridges that connect cells are already present in the basal *T. socialis*, which suggests an early origin for cytoplasmic bridges in the lineage.

In *V. carteri* and closely related taxa cytoplasmic bridges are important during embryogenesis but later the ECM replaces their role in cellular connectivity (Kirk and Nishii 2001). In other sections of *Volvox* cytoplasmic bridges are retained in the mature form. As yet little is known regarding what the molecular mechanisms are that underpin how cytoplasmic bridges are differently regulated throughout the lineage. Additional molecules that promote adhesion, such as metazoan cadherins, selectins and integrins (Niklas and Newman 2013), remain unknown in the volvocines. Tantalising evidence suggests that molecules with homology to fasciclin-domain containing metazoan cell adhesion molecules may be important for cell adhesion. A fasciclin-domain containing protein named algal-cell adhesion molecule (Algal-CAM1) has been identified in *V. carteri* and it may play a role in cell adhesion as well as inversion (Huber and Sumper 1994). Treating synchronised *V. carteri* cultures with antibodies targeting these proteins prior to embryogensis form misshapen ‘doughnut’ shaped blastomeres (Huber and Sumper 1994). Furthermore, genes that regulate the cell cycle, such as the retinoblastoma gene, may indirectly be important for incomplete cytokinesis (Hanschen, et al. 2016).

1.3.4.2 Inversion

Investigations into the molecular basis for inversion have largely focused on the kinesin *invA*, which if knocked-out creates an inversion negative (or *inversionless*) phenotype in *V. carteri* (Nishii, et al. 2003).

The mechanism by which *invA* acts is by localising to cytoplasmic bridges where it subsequently supposedly moves along cortical microtubules, creating torsion, thereby forcing cells to re-orientate around cytoplasmic bridges (Nishii, et al. 2003; Nishii and Miller 2010). The *C. reinhardtii* orthologue of *invA* - *IAR1* – performs the same function in *invA* (-) *V. carteri* (Kirk 2005). Currently, the function of this gene in unicellular and simple colonial taxa is unknown, however, in *V. carteri invA* is involved in inversion. There are two mechanical approaches for embryo inversion in the lineage such that the process differs between *V. carteri* (type A inversion) and *Volvox globator* (type B inversion) (Hohn and Hallmann 2011; Haas and Goldstein 2015; Hohn, et al. 2015). The genetic differences that explain these alternative approaches are unknown. Orthologues of *invA* are also known from species of *Gonium*, *Pandorina*, *Pleodorina* and *Eudorina*. Two additional genes known as *invB* and *invC* are not considered to be active in inversion but knockouts also results in inversion-less phenotypes (Coleman 2012). Both *invB* and *invC* are involved in membrane expansion/enlargement, which may be required for cells to move and re-orientate freely within the developing embryo.

As already mentioned, the *algal-CAM* gene may be important for inversion. When an antibody targeting the algal-CAM protein was applied later in *V. carteri* embryogenesis but prior to inversion an inversion-less (but otherwise normal phenotype) was observed (Huber and Sumper 1994). As per *invB* and *invC*, cell adhesion mediated by algal-CAMs may be a requirement of inversion in *V. carteri* but is unknown if it is an active participant in the process of inversion. Besides the role of InvA little is known about additional proteins actively involved in embryo inversion. Interest in identifying additional genes active in inversion was expressed by Kirk in 2005 (Kirk 2005) but the outcome of these efforts has not been communicated.

1.3.4.3 Basal Body Rotation

Although basal body proteins have been identified from the genomes of *C. reinhardtii, G. pectorale* and *V. carteri*, discriminating molecular characteristics have not been identified to explain basal bodies re-orientation in the multicellular volvocines.

1.3.4.4 Organismal polarity

The genetic underpinnings of organismal polarity remain unresolved. However, volvocine embryology has been well investigated in many taxa and to some extent the physical and mechanical aspects of how embryos develop are known. One candidate protein of interest is heat shock protein 70 (Hsp70), because Hsp70 is expressed in a gradient

during *V. carteri* asymmetrical division - with expression most abundant in the anterior of a developing blastomere (Cheng, et al. 2005). It would be of interest to determine if a similar expression pattern is found in other multicellular taxa without asymmetrical division, such as *Eudorina*. Asymmetrical division evolved later than organismal polarity but may be co-regulated. Various forms of organismal polarity are found throughout the lineage and it may thus be challenging to identify a unifying mechanism. However, organismal polarity may in fact not be programmed via a regulatory process but might instead be a random consequence of embryogenesis that occurs due to physical constraints.

1.3.4.5 Transformation of the Cell wall into an ECM and the Expansion of the ECM

A few molecular innovations have been associated with this development. The biochemical characteristics of volvocine cell walls and the extracellular matrix have been investigated (Goodenough and Heuser 1985; Kirk, et al. 1986; Woessner and Goodenough 1989; Ertl, et al. 1992; Woessner, et al. 1994; Godl, et al. 1995; Godl, et al. 1997; Sumper and Hallmann 1998; Hallmann 1999; Hallmann and Kirk 2000; Hallmann 2003). Furthermore, comparative genomic investigations of *C. reinhardtii*, *G. pectorale* and *V. carteri* have identified genes/proteins that associate with both the loss/modification of the outer cell wall as well as the expansion of the ECM, which is a characteristic feature of *V. carteri* where the ECM can constitute >95% of organismal volume (Coleman 2012).

The glycoproteins GP1, GP2 and GP3 are an integral component of the *C. reinhardtii* outer cell wall (Goodenough, et al. 1986; Adair, et al. 1987) and orthologues of the GP1 protein are lost in both *V. carteri* and *G. pectorale* (Prochnik, et al. 2010; Hanschen, et al. 2016). The loss of GP1 in *V. carteri* and *G. pectorale* might be associated with the modification of the outer cell wall. Genes encoding for pherophorin glycoproteins and matrix metalloproteases (*Volvox* Matrix Metalloproteases – VMPs- also known as MMPs or gametolysins) are greatly expanded in *V. carteri (*Prochnik, et al. 2010; Hanschen, et al. 2016). Using comparative genomic analyses, the expansion of these families has been the only large-scale expansion of gene families associated with morphological traits identified thus far. The pherophorins are hydroxy-proline rich glycoproteins similar in function to plant extensins or lectins, which function in the volvocines to bind simple carbohydrates and thereby act as scaffolds to support the gelatinous ECM (Goodenough and Heuser 1985; Godl, et al. 1995; Ender, et al. 2002). Pherophorins are characterised by possessing the Hallman pherophorin domain and may contain multiple copies of this domain (Godl, et al. 1995; Godl, et al. 1997). As part of sexual induction their expression has also been shown to be upregulated in *V. carteri* by sex inducers and by nitrogen starvation in *C. reinhardtii* (Sumper, et al. 1993; Amon, et al. 1998; Rodriguez, et al. 1999).

The number of genes coding for pherophorins is greatly expanded in *V. carteri* (n=78) relative to *C. reinhardtii* (n=35) and *G. pectorale* (n=31)

and thus their expansion is more appropriately associated with the expansion of the ECM than the transformation of the ECM into a cell wall. The matrix metalloproteases (MMPs) are another major constituent of proteins found in the ECM and have been identified in a number of volvocine taxa (Hallmann, et al. 2001). Metalloprotease proteins are characterised by comprising of a metalloprotease metal binding domain (for copper or zinc binding) and a proline rich domain (HR domain) (Heitzer and Hallmann 2002).

The substrates of VMPs have not been explored (Heitzer and Hallmann 2002) but have been shown to be involved in degrading the cell wall during gamete release and hence are also referred to a gametolysins (Goodenough and Heuser 1985). In metazoa MMPs play an important role in shaping the ECM through degradation of ECM components (Hallmann, et al. 2001). *Volvox* metalloprotease expression can also be triggered by wounding, which indicates a role in ECM remodelling response (Hallmann, et al. 2001). However, in *V. carteri* the timing of VMP expression appears to indicate that the dominant role of VMPs is to degrade pherophorins upon induction by sex-inducing pheromones (Hallmann, et al. 2001). This process of degradation may be part of a signal transduction cascade that activates the sexual reproductive pathway. Relative to *C. reinhardtii* and *G. pectorale*, the expansion of VMP coding genes is also an exclusive feature of *V. carteri* with 44 genes in *C. reinhardtii*, 36 genes in *G. pectorale* and 98 genes in *V. carteri*. It is intriguing that the ratio of MMPs

to pherophorins in each of the whole-genome sequenced organisms remains similar (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). Heitzer and Hallman (2002) speculated that specific MMPs might exist to target specific pherophorins, which themselves occur throughout different layers of the ECM. Genomic evidence alone is insufficient to test this hypothesis, but has not contradicted it. Experimental evidence that identifies the exact substrates of MMPs should be performed.

1.3.4.6 Genetic Modulation of Cell Number and the Modification of the Cell Cycle

The maximum number of divisions (N or maximum cell number) is a defining characteristic of multicellular volvocines that is used for the morphological and taxonomic classification in the lineage. The unicellular multiple-fission-division pathway where the varying size of the mother cell determines the varying number of divisions has been modified in the multicellular taxa into a programmed number of divisions (Kirk 2005). The assertion that the adaptation of the cell cycle was due to a genetic modification of the cell cycle regulatory programme is an important premise when considering the emergence of multicellularity in the volvocine lineage. An important aspect of regulating cell number is the regulation of cell size. Various challenges associated with control over cell size and number of divisions have been raised, such as: what do cells measure to determine the readiness for division (sizer molecule/s?), how

cells address the dilution of key regulatory cell cycle proteins throughout division cycles and how the minimum threshold cell size is determined (which may differ by cell type) (John 1984; Umen and Goodenough 2001; Li, et al. 2016)?

A number of models that attempt to explain the control of cell size have been posited, including those that propose sizer molecules, those that propose that the gradient of molecules from the nucleus to the outer wall determines readiness for division and those that theorise that the distance of centrosomes from the inner cell wall/membrane determines cell size as well as models that propose a mixture of factors or lineage specific models (Donnan and John 1983; Marguerat and Bahler 2012; Marshall, et al. 2012). Part of the challenge of obtaining a mechanism for understanding how cell size is controlled that is fundamental to all of life is the diversity of reproductive modes e.g. binary fission, budding and palintomy.

Efforts to uncover the genetic underpinnings of volvocine cell cycle modification have largely focused on what may be referred to as ‘core’ cell cycle genes. The term core cell cycle genes is not a strict definition but, more generally, such genes form part of most eukaryotic cell cycle pathways as key mediators. Core cell cycle genes include *cyclins*, *cyclin-dependent protein kinases* (*CDKs*), DNA damage checkpoint genes, certain transcription factors and genes such as the *retinoblastoma* (*RB*)

gene (Umen and Goodenough 2001; Bisova, et al. 2005; Cross and Umen 2015; Li, et al. 2016). Cyclins typically form part of complexes with other proteins, such as CDKs, and these complexes drive cells through the cycle. Cyclin-dependent kinases also form part of complexes and regulate the function of cyclins and other proteins through phosphorylation. In animals it is clear that in the presence of DNA damage various checkpoint genes inhibit cell cycle progression. The *RB* gene also plays a key role in supressing transition through the cell cycle. *Cyclins*, *CDKs*, components of the *RB* pathway as well as common eukaryotic cell cycle genes have been identified in the volvocine lineage (Umen and Goodenough 2001; Merchant, et al. 2007; Olson, et al. 2010; Prochnik, et al. 2010; Hanschen, et al. 2016; Li, et al. 2016).

To understand how the cell cycle of multicellular taxa may have been modified it is helpful to first understand the *C. reinhardtii* cell cycle. The proposed programme that defines the number of divisions in *C. reinhardtii* is based on a timer-sizer model where an intrinsic timing module (presumably related to environmental day/night cues) activates the process of accessing the state of readiness for undergoing cell division (Matsumura, et al. 2003; Kirk 2005; Cross and Umen 2015). The job of assessing the readiness for division falls upon sizer molecule/s, which assess readiness via assessing cell size. After each round of division, the readiness for division is reassessed, proceeding until a minimum threshold cell size is achieved, after which, division ceases and daughter cells are of

uniform size. With regard to palintomic multiple-fission-division programmes, *C. reinhardtii* is a highly useful model organism for studying the mechanics of cell division and the cell cycle. In *C. reinhardtii* the *retinoblastoma* (*MAT3*/*RB*) pathway and the *cyclin-dependent kinase G1* (*CDKG1*) have been identified as key determinants of cell cycle progression and cell size. The *MAT3* gene is a regulator at the G1-S checkpoint (and for entry into each palintomic division) (Umen and Goodenough 2001) while *CDKG1* plays a role in sensing cell size (a sizer molecule) at the S-M phase (Li, et al. 2016).

1.3.4.6.1 The retinoblastoma homologue

The RB protein is a key negative regulator of the cell cycle in animals and in photosynthetic eukaryotes (including the volvocines). In animals it performs this role through binding and thereby capturing elongation 2 transcription factor (E2F) protein/s, which when unbound to RB forms a complex with dimerisation protein (DP) and activates the transcription of cell cycle promoting genes (Muller, et al. 1994; Narasimha, et al. 2014). In animals, the binding of RB to E2F-DP is regulated by its state of phosphorylation, which is controlled by CDK proteins, whereas in *C. reinhardtii* RB remains in complex with E2F-DP and its dissociation from E2F-DP is not a requirement for cell cycle progression (Olson, et al. 2010). A possible explanation for this is that RB/MAT3 phosphorylation, which is most prevalent at the S-M phase, may alter the conformation of the MAT3-E2F-DP complex and in so doing initiates the transcription

activation capacity of the complex (Olson, et al. 2010). Other proteins such as chromatin remodelling enzymes or other transcription factors may also play a role in this process. Knockouts of *MAT3* in *C. reinhardtii* create a small cell phenotype because cells prematurely pass through the commitment phase of G1 into the S phase and progress through further rounds of division where they are not inhibited at the S-M phase (Umen and Goodenough 2001). Interestingly, knockout mutations in *DP* or *E2F* can rescue the wild-type phenotype of a *MAT3*-knockout. Loss of *RB* in *A. thaliana* is lethal (Ebel, et al. 2004; Zhao, et al. 2017), while in humans the loss of *RB* is tumourigenic (Araki, et al. 1991; Nakamura, et al. 1991; Cryns, et al. 1993; Hogg, et al. 1993; Linardopoulos, et al. 1993; Presti, et al. 1996).

The comparative genome analysis of *C. reinhardtii*, *G. pectorale* and *V. carteri* identified two key factors related to the RB protein that are associated with the evolution of multicellularity: (i) relative to *C. reinhardtii* the linker region between RB-A and RB-B domains is shortened in *G. pectorale* and *V. carteri* and (ii) 2 CDK phosphorylation targets were identified in the linker region between RB-B and the carboxy(C)-terminal domain of *C. reinhardtii* RB but were absent in *G. pectorale* and *V. carteri* (Hanschen, et al. 2016). The shortened linker region between RB-A and RB-B was suggested to be important for binding E2F and DP while the absence of identifiable phosphorylation targets between RB-B and the C-terminal domain was speculated to be important for targeting and

regulation by CDKs (Hanschen, et al. 2016). Transforming a *mat3* (-) deficient *C. reinhardtii* strain with the *G. pectorale rb* gene created a non-palmelloid colonial *C. reinhardtii* morphology that varied in cell number from 2-16 cells. The degree of variation in cell number was not reported (Hanschen, et al. 2016). Interest in the volvocine *rb* gene extends to a possible role for the gene in oogamous sexual reproduction (Kianianmomeni, et al. 2008). This is because: (i) the gene has migrated from an autosomal position in chromosome 6 of *C. reinhardtii* to reside within the mating-locus of *G. pectorale* and *V. carteri* (Ferris, et al. 2010; Hamaji, et al. 2016), (ii) its role in determining cell size (as evident from *C. reinhardtii mat3* mutants) and (iii) because gender-specific splice variants are present in *V. carteri* (Kianianmomeni, et al. 2008; Ferris, et al. 2010). However, recent evidence suggests that the gender-specific *Volvox* splice variants may not be of general importance for anisogamous/oogamous sexual reproduction because the evolution of anisogamy and/or oogamy predates variations in the *rb* gene structure (Hiraide, et al. 2013).

1.3.4.6.2 Cyclin dependent kinase G1

While *MAT3*/*RB* assumes the role of a commitment checkpoint gene between the G1 and S phase as well as subsequent S-M interphases of the cell cycle, the cyclin-dependent kinase *CDKG1* performs the role of linking the number of divisions to the mother cell size (Li, et al. 2016). Both molecules can be described as sizer molecules but differ in that the MAT3 protein checks the size of the cell entering division

only, while CDKG1 concentration correlates with the size of the initial mother cell and consequently the number of divisions.

In contrast to a *MAT3*/*RB* knockout, *C. reinhardtii CDKG1* mutants result in a large-cell phenotype and the large cell phenotype is not due to an increased growth phase or delays in cell cycle progression (Li, et al. 2016). Instead the increased cell size of *CDKG1*-mutant daughter progeny is due to a relaxation of cell size monitoring. The peak concentration of CDKG1 is at the G1-S phase and correlates (allometrically) with the mother cell size. From the peak concentration CDKG1 concentration diminishes with each successive round of division until no detectable CDKG1 can be measured and further cell divisions stop. The diminishing of CDKG1 concentration occurs through both dilution (a halving with each division) and degradation. Targets of CDKG1 include MAT3 and Cyclin D proteins (Li, et al. 2016). CDKG1 is not a ubiquitous eukaryotic cell-dependent kinase and was first identified in *C. reinhardtii* but is also found in *G. pectorale* and *V. carteri*. It is unknown if an alteration of *CDKG1* sequence, activity or degradation is important in the evolution of the genetic modulation of cell number in the multicellular volvocines, but it is an interesting candidate for further exploration.

1.3.4.6.3 Other cell cycle determining genes

Another gene known to be involved in cell size determination is *FA2* (*CNK2*), which is a *never in mitosis-related* (NIMA) kinase with

serine/threonine kinase activity and plays a role in size determination and the process of deflagellation (Mahjoub, et al. 2002). With regard to a role in cell cycle progression *FA2* differs from *CDKG1* in that *FA2* null mutants are delayed in their progression through the cell cycle and thereby result in a larger daughter cell phenotype (Mahjoub, et al. 2002). But as is the case of *CDKG1* mutants *FA2* mutants do not regulate commitment cell size. Curiously, RNAi knockdown of wild-type *FA2* produces a normal phenotype and thus its concentration is not tied to cell size in the manner that *CDKG1* is. A number of *C. reinhardtii* mutants with aberrant cell cycle patterns have been identified but for many the causative genetic loci are not known (e.g. *ISO1*, *ISO2* and *ISO3*) (Campbell, et al. 1995; Coleman 2012). The repertoire of cell cycle genes in the multicellular taxa is similar to that of *C. reinhardtii* (Bisova, et al. 2005). The most notable exception is the expansion of D1-type *cyclins* in *G. pectorale* and *V. carteri*. The cyclin D1 (*CYCD1*) gene is present as a single copy in *C. reinhardtii*, whereas *CYCD1* is present in four copies in both *G. pectorale* and *V. carteri* (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). In both *G. pectorale* and *V. carteri CYCD1* expansions are found as tandem array duplications. Compared with other *cyclin* genes *G. pectorale* D1 *cyclins* have elevated rates of non-synonymous substitutions over synonymous substitutions (dN/dS) (Hanschen, et al. 2016).

The finding of *CYCD1* expansion in *G. pectorale* suggests that the expansion of D1 cyclins correlates with the emergence of simple colonial

volvocine forms rather than with the development of more derived features in *V. carteri* such as germ-soma differentiation. The *D-cyclins* are known from a variety of organisms, including *C. reinhardtii* to function as subunit activators of CDKs that target RB proteins (Cross and Umen 2015; Li, et al. 2016). Other observations from genome studies identified to date include a novel cyclin (*CYCD5*) in *G. pectorale* as well as expansions in *G. pectorale* of *CYCA1* (2 copies versus 1 copy in *C. reinhardtii* and *V. carteri*) and *CYCAB1* (3 copies versus 1 copy in *C. reinhardtii* and *V. carteri*).

Based on current evidence, modification of the RB pathway has been put forward as a primary underlying mechanism for the evolution of multicellularity in the volvocine lineage (Hanschen, et al. 2016). This is based on: (i) the expansion of D1-type *cyclins*, which together with *CDKs* act on *RB*, (ii) multicellular induction from cloning *G. pectorale RB* into *MAT3* (-) *C. reinhardtii*, (ii) *RB* domain-linker modifications identified in *G. pectorale* and *V. carteri*, and (iv) the known molecular function of *MAT3*/*RB* in *C. reinhardtii* for cell cycle regulation. However, some questions of how the RB pathway causes multicellularity remain, these include:

(i) Although the D1-type cyclins are expanded in *G. pectorale* and *V. carteri* it is unknown if they are required for the evolution multicellularity.

After all, *RB Gonium* transformation into *C. reinhardtii* (with a single copy of *CYCD1*) was sufficient to produce colonies.

(ii) Do the expanded copies of D1-type cyclins perform alternative roles (neo-functionalisation or sub-functionalisation) or does the expansion affect dosage? Similarly, do *G. pectorale* and *V. carteri* require all 4 copies or will 3 or 2 suffice? As yet no experimental approaches have been performed to address these questions.

(iii) The mechanism by which the RB protein regulates the *Chlamydomonas* cell cycle and other volvocines is unknown. This is especially important because RB dissociation from E2F and DP is not a requirement for cell cycle progression. It remains to be determined if differential phosphorylation of RB and activation by other proteins explains how RB regulates cell cycle progression. If other proteins interact with RB-E2F-DP complex what are these proteins?

(iv) It is unknown if transforming *RB* genes from other volvocine algae into *C. reinhardtii* will create the same colonial phenotype.

(v) The *in vivo* phosphorylation of RB in multicellular volvocine has not been measured; it is unknown if a difference between informatically-identified phosphorylation targets between the RB-B and C-terminal domain is of importance for RB regulation.

There has been much focus in the volvocine lineage on core cell cycle genes. However, in animals, yeast and plants much is known about additional cell cycle regulatory genes such as, additional checkpoint genes, transcriptional regulators, DNA damage/repair genes and a myriad of signalling genes that directly or indirectly act upon the cell cycle. Little is known about these extended regulators in *Chlamydomonas* or indeed the multicellular taxa. Given the relative similarity of the core cell cycle genes across the lineage additional regulators of the cell cycle could be of interest for the evolution of volvocine multicellularity.

1.3.4.7 The Molecular Underpinnings of Germ and Soma Differentiation

The development of germ-soma differentiation in the volvocine lineage is of great interest to evolutionary developmental biologists as it marks the development of division of labour in the lineage, which is a hallmark of complex multicellularity (Knoll 2011). Of particular interest to the evolution of volvocine germ-soma differentiation is the *somatic regenerator* (*regA*) transcription factor as well as *regA*-related (*rls*) genes (Harper, et al. 1987; Kirk 1994; Kirk, et al. 1999; Duncan, et al. 2006). The role of *regA* in supressing the reproductive capacity of soma has prompted *regA* to be labelled as an 'altruistic' gene (Nedelcu and Michod 2006). The fate of developing cells as either germ or soma within a *V. carteri* embryo is determined by cell size (Kirk, et al. 1993), which varies because asymmetrical division.

Besides the *rls* genes, genes involved in controlling asymmetrical cell division have been identified (Kirk 1997; Miller and Kirk 1999; Kirk 2001). Kirk proposed a molecular model for how cell fate is determined during *V. carteri* embryogenesis, which integrates the roles of *regA* and genes involved in asymmetrical division (Figure 1.6) (Kirk 2001; Kirk and Kirk 2004). Work by Kirk and co-workers as well as earlier work (e.g. by Starr, Huskey or Griffin) determined a number of important molecular aspects of *V. carteri* germ-soma differentiation (Starr 1968, 1970; Sessoms and Huskey 1973; Huskey and Griffin 1979; Kirk, et al. 1987). The model proposed by Kirk is briefly described hereafter. *Volvox carteri* asexually reproduces over approximately 48hrs, during which, cell divisions (n=11-12) occur over a relatively short time period (~7hrs) and cell growth does not occur between divisions. The first 5 divisions are symmetrical, after which, the *gonidialess* (*glsA*) gene is turned on. The *glsA* gene encodes a zuotin-family J-domain chaperone, similar to heat-shock protein 40 (*hsp40*), which functions through interactions with other proteins such as Hsp70 and the mitotic spindle to facilitate asymmetrical cell division (Miller and Kirk 1999; Cheng, et al. 2003; Cheng, et al. 2005). Notably an orthologue of *glsA*, *GAR1*, is present in *C. reinhardtii* and transformation of *GAR1* into *glsA* (-) *V. carteri* mutants can rescue the wild-type phenotype (Cheng, et al. 2003).

The size of the cell progeny signal for alternative developmental pathways such that large cells form gonidia whereas small cells become

somatic. After asymmetrical division gonidial-initials (cells fated to become gonidia) undergo a limited number of divisions (n=2-3), whereas somatic-initials continue to divide until the minimum somatic cell size is achieved (n=2-7). This difference in the number of divisions between gonidial- and somatic-initials is under the control of 3 genes known as *late gonidia* (*lag*) genes. The *lag* genes are expressed in gonidial-initials and turned-off in soma. The *lag* genes have not been identified in the genome of *V. carteri* and further functional analyses are required. After division the *regA* transcription factor is expressed in somatic cells and is not expressed in gonidial cells (Kirk, et al. 1999). The expression of *regA* in somatic cells supresses their reproductive capacity and after juveniles hatch somatic cells ultimately enter senescence and die.

The prevention of somatic cell reproduction and ultimately somatic cell death is why *regA* has been described metaphorically as an altruistic gene – it enforces altruistic behaviour in soma (Nedelcu and Michod 2006). The hypothesised mechanism by which *regA* halts reproduction in soma is via the suppression of chloroplast biogenesis, which stops cell growth and therefore cell division (Tam and Kirk 1991a; Kirk, et al. 1999). This working hypothesis is based on evidence that *regA* represses the transcription of nuclear genes coding for chloroplast proteins.

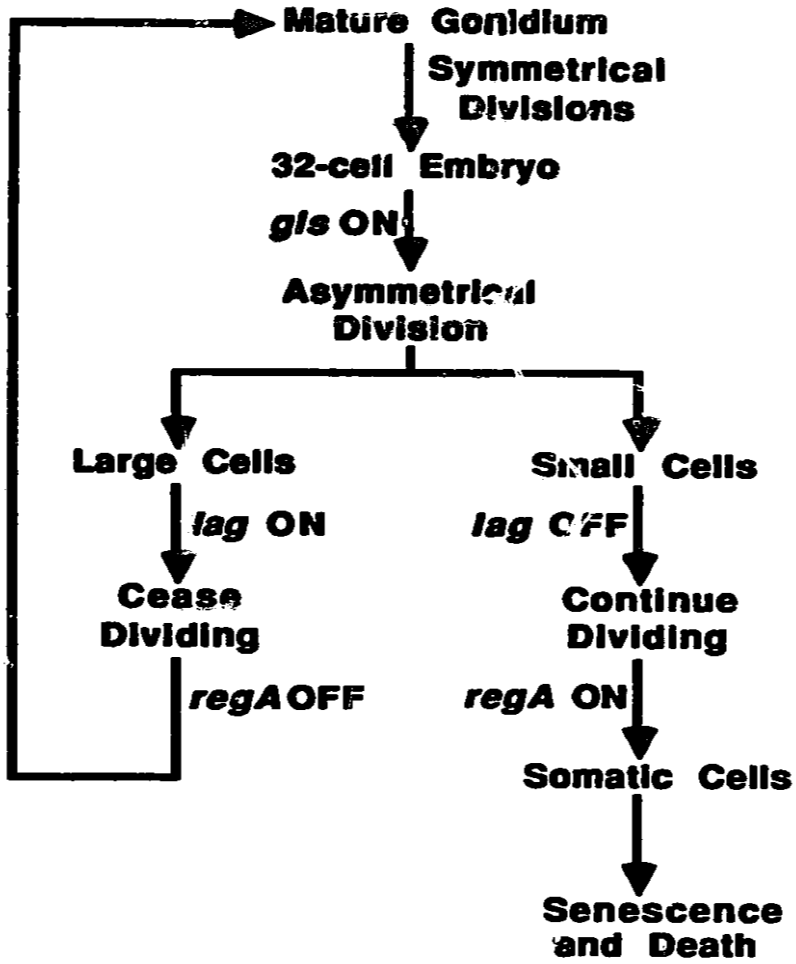


Figure 1.6: A working model of the Genetic underpinnings of Volvox germ-soma differentiation. Image taken from (Kirk and Kirk 2004).

RegA-related genes have been identified in the genomes of *C. reinhardtii*, *G. pectorale* and *V. carteri*, however, *regA* is absent in *C. reinhardtii*, and *G. pectorale* (Duncan, et al. 2007; Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). RegA-related proteins are highly divergent with only the core SAND domain or VARL region conserved amongst these proteins (Duncan, et al. 2007). SAND domains are typically limited to certain eukaryotic chromatin-dependent

transcription factors. *RegA*-related genes number 12, 8 and 13 respectively in *C. reinhardtii*, *G. pectorale* and *V. carteri*. Phylogenies of RLS proteins (using the VARL domain with or without including the N-terminal extension) show that *Volvox* spp. RegA is most closely related to the Rls proteins: RlsA, RlsB, RlsC and RlsN (Duncan, et al. 2007; Hanschen, et al. 2014). The cluster of *regA*, *rlsA*, *rlsB* and *rlsC* are labelled as the *regA* cluster and in organisms investigated to date they lie in close genomic proximity (Hanschen, et al. 2014; Grochau-Wright, et al. 2017). Furthermore, *rlsD* is closely related to *regA* cluster genes and is found in close proximity to the *regA* cluster (Prochnik, et al. 2010; Hanschen, et al. 2014; Grochau-Wright, et al. 2017). The current model suggests that *regA* genes are paralogs of *rlsD* (Hanschen, et al. 2014; Grochau-Wright, et al. 2017). No orthologues of *regA* cluster genes are found in *C. reinhardtii*, and *G. pectorale* (Merchant, et al. 2007; Hanschen, et al. 2016). However, *C. reinhardtii*/*G. pectorale RLS1* are probably orthologues of *rlsD* (Duncan, et al. 2007; Hanschen, et al. 2014). Orthologues of the *regA* cluster genes (including *regA*) have been identified in *Yamagishiella unicocca*, *V. carteri f. nagariensis*, *V. carteri f. kawasakiensis*, *V. carteri f. weismannia*, *V. gigas and V. ferrisii*, *Pandorina morum* and *Pleodorina californica* (Grochau-Wright, et al. 2017).

The current model of *regA* evolutionary history is that an *RLS1*/*rlsD*-like ancestral gene gave rise to the *regA* cluster through a series of tandem duplications that took place near the origin of the Volvocaceae

family (Hanschen, et al. 2014; Grochau-Wright, et al. 2017). It is interesting that *regA* cluster genes are present in taxa such as *P. morum*, which does not possess germ-soma differentiation. The origin of the *regA* cluster near the origin of the Volvocaceae potentially explains how germ-soma differentiation could have evolved multiple times in the Volvocaceae. However, it is unknown if the same molecular mechanisms control the various forms of Volvocacean germ-soma differentiation. For example, taxonomic considerations suggest that a similar mechanism is not likely to define germ-soma differentiation in *A. gubernaculifera*, which is a member of the Goniaceae not the Volvocaceae. Currently, it is unknown if *regA*-cluster genes besides *regA* are involved in germ-soma differentiation and indeed, more generally, what functions *rls* genes perform. It is clear, however, that the function of *regA* has been co-opted from an ancestral function and hence interest in *regA* has also driven interest into the ancestral role of *rls* genes. The role of *RLS* genes in unicellular organisms is intriguing because the hypothesised role of *regA* is to halt cell growth through the inhibition of chloroplast growth. This is an advantageous function in the multicellular volvocines, however, when viewed from a unicellular perspective this function is likely to reduce fecundity.

However, expression profiling of *C. reinhardtii RLS1* shows that *RLS1* expression is induced in the dark, which might yield a net improvement in resource allocation if chloroplast activity is suppressed in the dark (Nedelcu and Michod 2006). Findings indicate that *RLS1* is

upregulated under conditions of environmental stress where reproduction and growth may be of lessor survival benefit than maintaining viability (Nedelcu and Michod 2006). Together, these data suggest a plausible role for *RLS* gene function in unicellular chlorophytes. There remain many unanswered questions regarding the molecular underpinnings of germ and soma differentiation in the volvocine algae. This remains a highly active area of volvocine research for which various molecular experiments currently underway are expected to generate a more complete picture, for example, recent analyses have shown large-scale differences in gene expression profiles between germ and soma. Germ cells gene expression is enriched in genes of more ancient origin while somatic cells express more genes related to motility and metabolism (Matt and Umen 2018).

1.3.5 Genomic Developments in the Volvocine Lineage

Whole genome sequences of *Chlamydomonas reinhardtii*, *G. pectorale* and *Volvox carteri f. nagariensis* have been generated (Prochnik, et al. 2010) and their assembly statistics are presented in table 1.1. Perhaps the most noteworthy finding of volvocine comparative genomics is the similarity of protein-coding potential (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). Across the phylogenetic spectrum of life a general trend is that complex organisms tend to have larger genomes than simpler organisms (Lynch and Conery 2003). However, many well-known exceptions defy this trend – such as *Amoebae* with 100gb+ genomes - and these exceptions have led to the

term ‘the C value paradox’ (the C-value is a measure of genome size) (Cavalier-Smith 1978). The same paradox is often observed in total gene number: for example, the unicellular *C. reinhardtii* has a total gene number similar to that of humans. Simply stated the C-paradox is the frequently observed paradoxical relationship between genes/genomes with organismal complexity. Despite the difference in complexity between the volvocine species, the genome of *V. carteri* is only 20 megabases (mb) larger than *C. reinhardtii* (and this is largely due to increased repeat expansion not an increase in complex ‘high-entropy’ coding-sequence); *V. carteri* and *C. reinhardtii* have a similar number of genes and the organisms share most protein domains. The *G. pectorale* genome assembly is larger than that of *V. carteri* and again the coding content is similar to that of *C. reinhardtii*. Between *V. carteri* and *C. reinhardtii* only a limited number of domains and genes are expanded in either organism. Expanded gene families in *V. carteri* include ECM genes, *cycD1* genes and histone genes (Prochnik, et al. 2010). In fact, when comparing the multicellular volvocines to *C. reinhardtii* - gene family and protein domain loss or reductions are more frequent than gains (Prochnik, et al. 2010; Hanschen, et al. 2016). The majority of morphological developments in the lineage have been attributed to genes for which homologues are found in *C. reinhardtii*, these include *GAR1*/*glsA*, *IAR1*/*invA*, *MAT3*/*rb* and *RLS*/*regA* (Olson and Nedelcu 2016).

As such, the ‘co-option’ of ancestral genes for new functions rather than gene or domain novelty appears to have been the major means for morphological development in this lineage (Olson and Nedelcu 2016). By comparison, the formation of novel protein domains and gene has been identified as important for the development of metazoan multicellularity (Putnam, et al. 2007; Srivastava, et al. 2010; Moroz, et al. 2014) and yet it appears that in the volvocine lineage these developments were limited. It is most likely that the large gap observed in coding potential between metazoa and their unicellular relatives is due to missing extant simple multicellular lineages that are more deeply rooted than surviving metazoa.

Table 1.1: Volvocine genome assembly statistics. Assembly statistics were calculated using Quast (Gurevich, et al. 2013) and from reported values.

Organism and Assembly	Contig N50 (kb)	Scaffold N50 (mb)	Number of scaffolds	Total Assembly size	Number of protein coding loci
Chlamydomonas reinhardtii v5	90. 6	7.78	54	111.1	17 737
Gonium pectorale v1	16.3	1.27	2 373	148.8	17 984
Volvox carteri v1	43.9	1.49	1 265	137.8	15 669
Volvox carteri v2	85.2	2.6	434	131.8	14 971

1.3.5.1 Repetitive Elements and Synteny

A large portion of the genomic size difference between *C. reinhardtii* and *V. carteri* was ascribed to the expansion of repeat regions (Prochnik, et al. 2010). The *V. carteri* assembly includes roughly 5% extra

repetitive elements compared to *C. reinhardtii*. The various classes of repetitive elements are differentially distributed between these organisms, for example; long-terminal repeat (*ltr*) elements are nearly 5 times more common in *V. carteri* than in *C. reinhardtii*, long interspersed elements (*lines*) are almost twice as frequent in *C. reinhardtii* as in *V. carteri* and short interspersed elements are twice as frequent in *V. carteri* than in *C. reinhardtii* (Prochnik, et al. 2010). The transposable element known as *Jordan* (Miller, et al. 1993) was not identified in *C. reinhardtii* whereas it is present in *V. carteri* (Prochnik, et al. 2010). This element was possibly acquired via viral transfection. Transposable elements (TEs) have been identified as having a useful and perhaps under-appreciated role in genome evolution (as reviewed by (Oliver and Greene 2009)). Transposable elements have been identified to play a role in the generation of new coding regions as well as in the creation of other important non-coding regions. In certain instances TEs can have an important role in gene regulation by modulating gene expression or the expression of micro-RNAs (Oliver and Greene 2009). The role of transposable elements, and non-coding DNA in general, in the transition to multicellularity in the volvocine family has not been adequately addressed and yet this may be of interest given that these elements do differ substantially between *V. carteri* and *C. reinhardtii*. The forces responsible for genomic expansion/contraction within volvocine algae - and eukaryotes in general - are multifaceted and complex (Cavalier-Smith 1978; Brosius 1999; Lynch and Conery 2003; Cavalier-Smith 2005).

Insights into the mechanisms for genome expansion in the volvocine algae have largely originated from studies of organelle genome expansion, and evidence from studies of organelles suggest that non-adaptive processes have been pivotal in genome expansion. Volvocine organelle genome complexity appears to scale positively with organismal complexity (Smith, et al. 2013), whereby colonial and multicellular species have larger, more repeat-rich organelle DNAs than unicellular species. It has been hypothesized that this trend is a consequence of smaller effective population sizes (N_e) and/or lower organelle mutation rates (μ) in multicellular versus unicellular volvocines (Smith and Lee 2010; Smith, et al. 2013). Similar causes may be important for nuclear genome expansion of the multicellular algae, although this is less well studied. The increase in multicellular nuclear genome size is likely to be due to non-adaptive processes, but it is possible that the effect of genome level differences such as chromosomal re-arrangements or chromosomal availability remain important for explaining differences in morphology. An important but often neglected consideration for phenotypic evolution is the role of large genetic changes or syntenic rearrangements. The evolution of cancer phenotypes is associated with substantial changes in genome organisation such as changes in chromosome copy number, translocations or other genetic exchanges between chromosomes (Heng, Bremer, et al. 2006; Heng, Liu, et al. 2006; Heng, Stevens, et al. 2006; Heng 2007; Gisselsson, et al. 2008; Heng, et al. 2008).

This form of genetic alteration is considered by some to be critical for speciation and related morphological developments (Heng 2009). In the volvocine lineage the preservation of synteny is low (Figure 1.7). Conserved synteny between *C. reinhardtii* and *V. carteri* is comparable with humans and chickens (Prochnik, et al. 2010). Thus, although the coding potential of volvocine algae is similar, genome organisation has been highly modified. These differences have mostly been overlooked when considering the genomic underpinnings of volvocine evolution. Due to technological innovations the organisation of genomes is currently becoming more thoroughly investigated (Chen, et al. 2013; Onmus-Leone, et al. 2013; Zhang, et al. 2015; Mariano, et al. 2016; Vij, et al. 2016). It is now known that the accessibility of genomic regions for transcription factors and other regulatory elements is of critical importance for phenotypic determination (Belsky, et al. 2015; Sequeira-Mendes and Gutierrez 2016). For example, recent evidence shows that genome availability is more accurate for classifying human B-cell subtypes than RNA expression (Corces, et al. 2016). Other factors such as the 3-dimensional configuration of chromosomes and organisation within the nucleus may also be of critical importance (Sequeira-Mendes and Gutierrez 2016). This aspect of genome biology is still in its infancy and is entirely unexplored in the volvocine lineage. It is plausible that changes in the function of various co-opted volvocine genes are the result of syntenic reorganisation rather than due to overt changes in amino acid sequences.

This hypothesis requires further investigation and is only possible when near-complete genomes are available. However, in favour of this hypothesis is the fact that the repertoire of volvocine coding potential is similar but syntenic rearrangements are extensive. Besides genomic studies evidence from simpler cytogenetic screens estimate volvocine chromosome number to range from 4 to 17 (Coleman 2012). Such differences are congruent with a view that espouses the important role of genome organisation (Heng 2009).

Besides large-scale syntenic differences, much evidence from recent literature suggests an important role for non-coding DNA in the regulation of gene function (Drosophila 12 Genomes, et al. 2007; Hubisz and Pollard 2014) and in this regard the volvocine algae differ substantially. Besides the coding potential of various volvocine algae, differences in genome organisation and non-coding genome evolution should be a focus of investigation. To this end the volvocine lineage may be able to provide important insights into genome evolution during a major evolutionary transition in complexity.

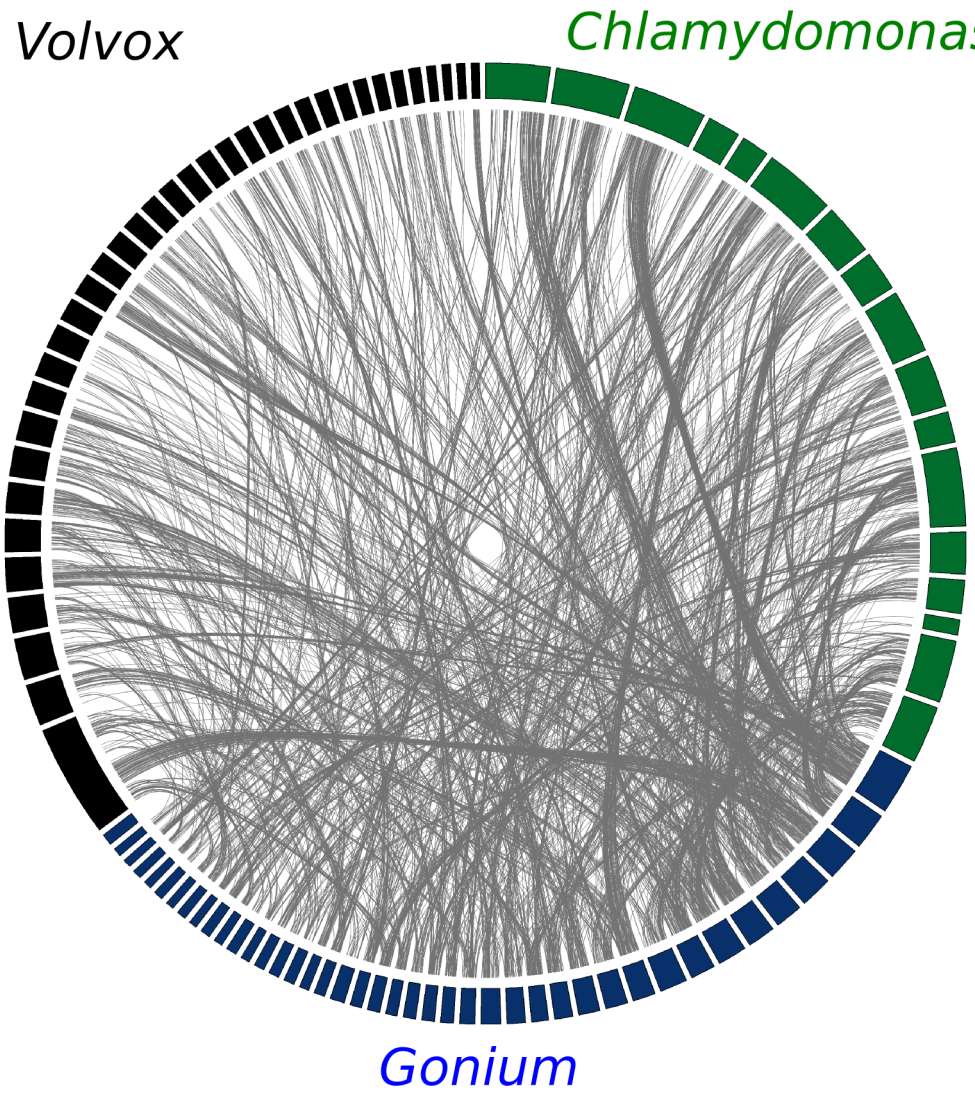


Figure 1.7: Synteny of *Gonium* with *Chlamydomonas* and *Volvox*. *Chlamydomonas* blocks in green are chromosome while blocks in *Gonium* and *Volvox* are scaffolds. Synteny between *Volvox* and *Chlamydomonas* is not shown in this image. Synteny is present, however, extensive rearrangements are also present. Gene-level synteny between *Volvox* and *Chlamydomonas* is comparable with Human and Chicken synteny (Prochnik, et al. 2010). Image taken with permission from (Hanschen, et al. 2016).

1.3.6 Sexual Reproduction and The Mating Locus

Besides the development of multicellularity, the volvocine lineage serves as model for studying the development of sexual dimorphism. Chapter 4 describes volvocine mating and the genetics thereof in greater detail. Sexual reproduction in the simpler volvocines such as *C. reinhardtii*, *T. socialis* and *G. pectorale* occurs through the formation of equal sized gametes (isogamy) whereas in more complex genera such as *Eudorina*, *Pleodorina* and *Volvox* gametes are of unequal size (Figure 1.8) (Coleman 2012; Nozaki, et al. 2014). In *Pleodorina* and *Eudorina* the gametes are of unequal size but both are motile (anisogamy) whereas in most *Volvox* species the female gamete is non-motile (oogamy). Genders are labelled differently between isogamous and anisogamous/oogamous types such that in *C. reinhardtii* the gender types are known as mating-type plus (*MT*+) and mating-type minus (*MT*-) and in *V. carteri* genders are known as male and female.

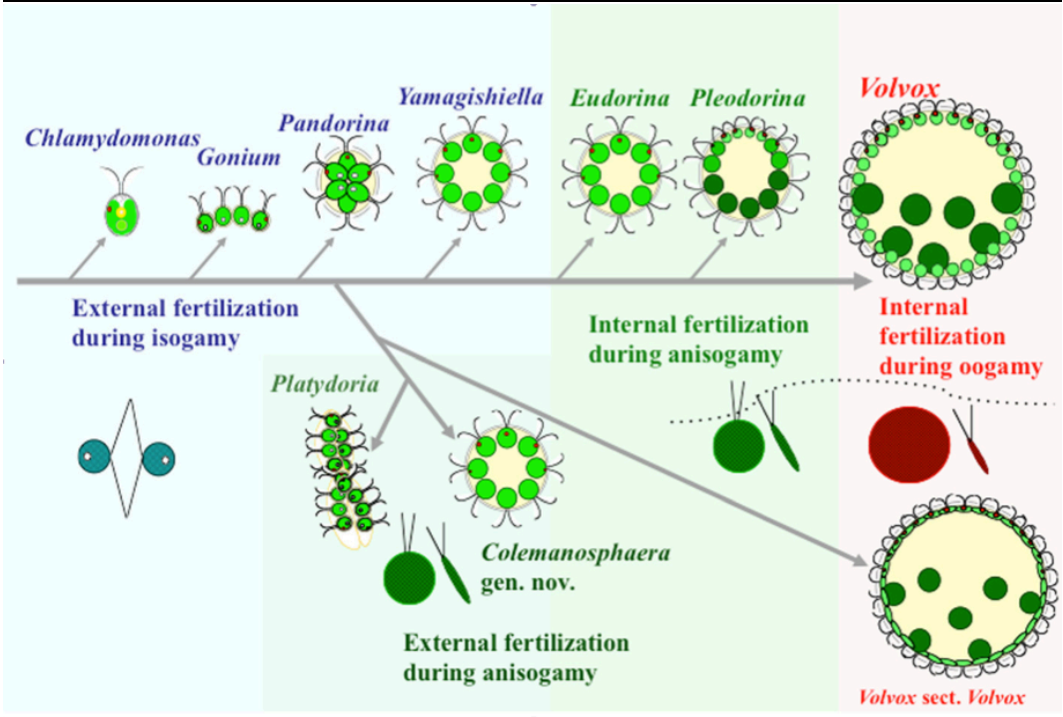


Figure 1.8: Types of sexual reproduction in the volvocine lineage.

From left to right are isogamy (equal-sized gametes), anisogamy (sexual dimorphism where both gametes are motile) and oogamy (sexual dimorphism where only sperm is motile). This figure presents alternative forms of anisogamy (interior versus exterior fertilisation). Of note is the independent evolution of oogamy in two clades of *Volvox*. Image from (Nozaki, et al. 2014).

In ideal conditions such as in nutrient rich environments volvocine algae reproduce vegetatively (asexually). In *Chlamydomonas* N-deprivation induces the sexual life (Amon, et al. 1998; Rodriguez, et al. 1999) whereas in *Volvox* the sexual life cycle is induced by species-specific sex pheromones produced by spermatozoa or by somatic cells

under heat stress (Jaenicke, et al. 1993; Sumper, et al. 1993). The sex pheromones are diffusible glycoproteins, similar to ECM components (Sumper, et al. 1993; Godl, et al. 1997; Amon, et al. 1998).

Response to sex-inducers is thought to initiate a cascade of signalling pathways that result in the formation of gametes from gonidial cells, the breakdown of the ECM and the release of gametes (Ertl, et al. 1992; Godl, et al. 1995; Hallmann, et al. 1998; Ender, et al. 1999).

A sequence region of low recombination governs gender determination in the volvocine lineage. Volvocine mating loci for both genders have been sequenced for *C. reinhardtii*, *G. pectorale* and *V. carteri* (Ferris and Goodenough 1994; Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). In all three species investigated, the region shows low rates of recombination and to an extent synteny is conserved. The region ranges in size from 200-300kb in *C. reinhardtii* to 360-500kb in *G. pectorale* and 1.13 to 1.15mb in *V. carteri*. The mating loci of *C. reinhardtii*, *G. pectorale* and more especially *V. carteri* are repeat rich and sequencing of these loci required sequencing of bacterial artificial chromosome (BAC) clones. The mating-locus is the “master-regulator” of sex in heterothallic (genetically determined gender) volvocines and it acts as a form of sex chromosome (Umen 2011). Genes within the mating-loci of *C. reinhardtii*, *G. pectorale* and *V. carteri* have been identified that are unique to either gender. The most gender limited gene is the putative

RWP-RK transcription factor known as the minus dominance gene (*MID*), which is found exclusively in the male or *MT*- haplotypes (Ferris and Goodenough 1997; Nozaki, et al. 2006; Hamaji, et al. 2009; Hamaji, Ferris, et al. 2013). To date *mid* has been identified in a few male or *MT*- volvocine taxa, including *C. reinhardtii*, *Chlamydomonas incerta*, a number *Gonium* species, *P. starrii*, *V. africanus* and *V. carteri*.

The *mid* gene is the dominant determinant of maleness in the lineage and may be described as a sex-determining gene. Other gender-specific genes include *MTD1* (*MT*- or male), *FUS1* (*MT*+), *EZY2*, *SRL1* as well as a number of genes of unknown function in *C. reinhardtii*, *G. pectorale* and *V. carteri* (Armbrust, et al. 1993; Ferris, et al. 1996; Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). In addition, the *FUS1* gene, which is likely to be involved in gamete fusion, and the *MTD1* gene is not present in either *V. carteri* haplotypes. Another topic of interest is that homothallic (the genetic capability to produce either male or female offspring) taxa are found throughout the lineage, including: the basal *T. socialis* and a number of species of *Gonium*, *Eudorina*, *Pleodorina* and *Volvox* (Coleman 2012). To date, mating loci are only available for heterothallic volvocines. The genetic differences between heterothallic and homothallic taxa are unknown.

1.3.7 Volvocine Organelle Genomes

Whole mitochondrial and plastid DNA (mtDNA and ptDNA) sequence data exist for a wide range of volvocine taxa, including *C. reinhardtii* (Boer, et al. 1985; Maul, et al. 2002), *G. pectorale* (Hamaji, Smith, et al. 2013), *Y. unicocca*, *Eudorina* sp. (Hamaji, et al. 2017), *P. starrii* 2013 (Smith, et al. 2013), and *V. carteri* (Smith and Lee 2009), as well as for several non-volvocine chlamydomonadalean species (Smith, et al. 2010; Del Vasto, et al. 2015). Volvocine organelle genome data have revealed a tremendous diversity in organelle genome architecture. For example, some volvocine algae, including *C. reinhardtii* and the multicellular *Pandorina morum* and *Y. Unicocca* have linear mtDNAs with defined telomeres (Moore and Coleman 1989; Vahrenholz, et al. 1993; Hamaji, et al. 2017), whereas *G. pectorale*, *Eudorina* sp (circular and linear), *P. starrii*, and *V. carteri* have circular-mapping mtDNAs (Smith and Lee 2009; Hamaji, Smith, et al. 2013; Smith, et al. 2013; Hamaji, et al. 2017), which in certain instances contain large numbers of mitochondrial repeats (Smith and Lee 2009). Similarly, the *V. carteri* ptDNA is dense with repeats and at ~525 kilobases (kb) is more than twice the size of that from *C. reinhardtii* (~204 kb), and is amongst the largest plastid genomes ever observed. Conversely, Reinhardtinia species from the non-photosynthetic genus *Polytomella* have lost their plastid genomes entirely (Smith and Lee 2014). The organelle genomes of *T. socialis* are discussed in greater detail in Chapter 2 (Featherston, et al. 2016). Subsequent to publication of the *T. socialis* organelle genomes findings from the

mitochondrial genomes of *Y. unicocca* and *Eudorina* sp have been published (Hamaji, et al. 2017). Together these results describe a complex scenario of multiple independent transitions between linear and circular mtDNA genome conformation in the volvocine lineage.

1.3.8 *Tetrabaena Socialis*

Tetrabaena socialis (Figure 1.9) is the 'type-species' for the Tetrabaenaceae family of volvocines. The other genus in the family is *Basichlamys*. The Tetrabaenaceae are sister to the more morphologically complex Goniaceae and the Volvocaceae. Until recently a morphological description of the 4-celled *T. socialis* was outstanding and *T. socialis* was initially classified as *Gonium sociale*. Later phylogenetic methods identified the Tetrabaenaceae as a separate family from the Goniaceae and the Volvocaceae. Recently, Arakaki et al (2013) provided an ultrastructural morphological description of *T. socialis*, associated morphological traits with Kirk's 12 steps and thereby identified 5 of the first 6 of Kirk's steps to be present in the basal *T. socialis*. The following of Kirk's steps were identified: 'incomplete cytokinesis', 'rotation of basal bodies', 'establishment of organismic polarity', 'transformation of cell walls into an ECM' and 'genetic modulation of cell number'. Arakaki et al (2013) made other interesting observations in their description of *T. socialis*. These include that the angle of orientation of *T. socialis* microtubule rootlets are intermediate between those of *C. reinhardtii* and *G. pectorale* and that, as per *C. reinhardtii*, *T. socialis* eyespots are 2-3 layered, which is unlike

more complex multicellular taxa that possess between 4-8 layered eyespots. However, the shape of *T. socialis* eyespots was similar (concave anterior facing) to those of more complex volvocines and different to those of *C. reinhardtii*. Cytoplasmic bridges connecting *T. socialis* cells were confirmed using transition electron microscopy (TEM) and their formation within dividing *T. socialis* embryos were observed. Morphological and phylogenetic features place *T. socialis* as amongst the simplest of extant 'integrated multicellular organisms' and key for understanding basal developments in the volvocine lineage (Arakaki, et al. 2013). As identified by Arakaki et al (2013) a number of morphology features of *T. socialis* are intermediate between *C. reinhardtii* and more derived multicellular taxa.

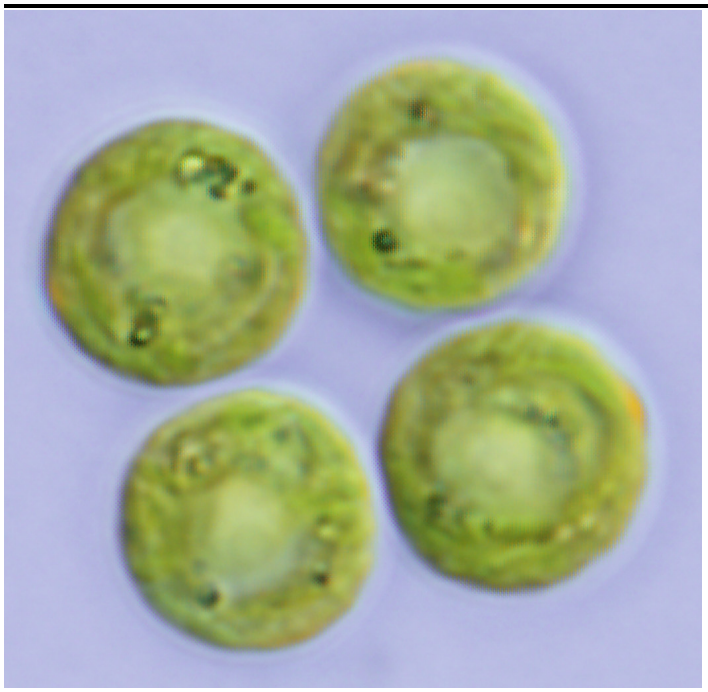


Figure 1.9: A light microscope image of *Tetrabaena socialis.*

Image kindly provided by Dr Yoko Arakaki. The extracellular matrix surrounding each cell is faintly visible.

1.3.9 A Model for the Evolution of *Tetrabaena Socialis*

Recently, a gradual model for the evolution of simple colonial volvocine multicellularity has been put forth (Maliet, et al. 2015). The model employed *T. socialis* as a proxy for the simplest form of multicellularity in the lineage. The model focuses on modifications in cell cycle timing (heterochrony) and emphasises the role of cell cycle modulation for the formation of a simple multicellular organism. A key premise of this model is that the difference between the cell cycle of an organism like *C. reinhardtii* and *T. socialis* can be described by a

difference in the time that the products of cell division remain connected after cell division (Maliet, et al. 2015). After cell division in *C. reinhardtii*, cell separation/cleavage follows shortly after division and is complete, while in *T. socialis* the time between division and separation is extended. This time difference allows for cell growth to occur in *T. socialis* while cells are connected but not in *C. reinhardtii*. Thus, *C. reinhardtii* exists in a temporary multicellular state and the extension of this state can result in an organism that remains as a single entity during the growth phase of the cell cycle. The model was presented mathematically but is essentially intuitively simple and provides an explanation for how a gradual process could give rise to simple multicellularity. In this model, selection favours gradual increases in the time that cells remain connected after cell division until the transition from single-celled to group living became fixed.

The model is in accordance with typical neo-Darwinian thinking, which favours gradual adaptation over the dramatic appearance of statistically improbable complex forms, which Stephan Gould caricatured as 'Hopeful Monsters'. Kirk's conclusion that the first 5 steps of the transition to the *Volvox* form possibly appeared simultaneously is an example of a hopeful monster (Kirk 2005). Kirk's position cannot be ruled out, however, the model by Maliet et al (2015) at the very least provides a gradual framework for cell cycle modulation, which in turn favours the gradual acquisition of additional traits. Although the volvocine algae offer the greatest diversity of extant taxa of varying degrees of morphological

complexity, forms with intermediate properties may be extinct and their absence may bias efforts to model the transition. Such intermediate species may not have survived since the emergence of the multicellularity (~230MYA) and, speculatively, such intermediate forms may only have possessed contingent fitness benefits in certain environments e.g. where acute selection for size existed due to predation.

1.4 Conclusions

The development of multicellularity has occurred numerous times in life's history and the understanding of how unicells may have cooperated to form complex multicellular organisms is of importance for academic reasons as well as for a better understanding of cooperating systems. The volvocine lineage is a useful model for studying the evolution of multicellularity, as there are many extant members with varying degrees of complexity. Genomes for *Chlamydomonas reinhardtii, Gonium pectorale* and *Volvox carteri* are available for study. Genome sequencing of extant organisms within the volvocine lineage will add further resolution towards understanding the genome-wide changes that have occurred in the development of multicellularity within this lineage. It is anticipated that the genome of *T. socialis* will contribute towards understanding the genetic underpinnings of simple multicellularity and the emergence of early developments in the volvocine lineage.

2 ORGANELLE GENOME EVOLUTION IN THE VOLVOCINES

2.1 Summary and Background of Manuscripts

This chapter focuses on organelle genomes in the volvocine lineage. Three publications are associated with this chapter related to volvocine organelle genomes. The manuscript by Featherston et al (2016) was initiated and written as part of this thesis while the other manuscripts were initiated and written by research collaborators. In Smith et al (2013) organelle genomes of *Pleodorina starrii* were sequenced using data that initially formed part of this thesis. In this manuscript it was shown that the size of volvocine organelle genomes was inversely correlated with effective population size (N_e) * mutation rate (**μ**). Furthermore, $N_e\mu$ correlated inversely with organismal complexity, suggesting that increases in organismal complexity resulted in a reduction in $N_e\mu$, which consequently resulted in inflated organelle genome structures. The implication of these findings is that population genetic effects determined organelle genome size due to neutral evolution. With increased organismal complexity comes smaller effective population size thereby diminishing the power of selection to remove potentially hazardous non-coding DNA.

In Del Vasto et al (2015), the organelle genomes of *Dunaliella viridis* and a previously un-sequenced strain of *Dunaliella salina* were

sequenced *de novo*. The Dunaliellaceae are members of Chlamydomonadales order of chlorophytes, which is the same order that the volvocines are part of. Genetic diversity in *Dunaliella* spp. was explored and was found to differ substantially between plastid and mtDNA genomes. The expansion of *Dunaliella* spp. organelle genomes was discussed in the context of population genetics. In Featherston et al (2016) the organelle genomes of *Tetrabeana socialis* were described. The organelle genomes of *T. socialis* were unexpectedly large and did not correlate with organismal complexity as per Smith et al (2013). The plastid genome of *T. socialis* at >405kb is massive. Largely based on previous analyses that show a relationship between volvocine organelle genome size and $N_e\mu$, the manuscript discusses possible causes for increased organelle genome size. Both linear and circular mitochondrial (mtDNA) genomes are known from the Reinhardtii-family. Phylogenetic reconstruction suggests that multiple transitions between linear and circular architecture may have occurred. In Featherston et al (2016) the mtDNA genome of *T. socialis* was found to be circular mapping, which suggests that a circular mapping mtDNA genome was present near the origin of multicellularity.

David R Smith, Takashi Hamaji, Bradley J.S.C Olson, Pierre M Durand, Patrick Ferris, Richard E Michod, Hisayoshi Nozaki, Jonathan Featherston and Patrick J Keeling. (2013). Organelle genome complexity scales positively

with organism size in volvocine green algae. *Molecular Biology and Evolution.* 30(4):793-797.

Michael Del Vasto, Francisco Figueroa-Martinez, Jonathan Featherston, Marilea Gonzalez, Adrian Reyes-prieto, Pierre M Durand and David Roy Smith. (2015). Massive and widespread organelle genomic expansion in the green algal genus Dunaliella*. Genome Biology and Evolution*. 7(3):656-663.

Jonathan Featherston, Yoko Arakaki, Hisayoshi Nozaki, Pierre M Durand and David R Smith (2016). Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in the volvocine green algae. *European Journal of Phycology*. 51(4):369-377

3 THE NUCLEAR GENOME OF *TETRABAENA SOCIALIS*

3.1 Summary of Manuscripts

The manuscript by Featherston et al (2018) is the central component of this thesis. In this publication the nuclear genome of *Tetrabaena socialis* is described and findings from comparative genomic analyses with other volvocines were interpreted in relation to the origin of simple colonial multicellularity. A genetic programme for the control of cell number was one of the earliest traits to develop in the multicellular volvocines and therefore findings described in the manuscript were interpreted in the context of the evolution of this trait. A conservative set of gene families were identified that associate with the evolution of multicellularity. Furthermore, gene families with expanded membership were identified. Overall, these gene families were rich in developmental related functions, although no functional annotations could be assigned to a significant percentage of gained gene families because they were not homologous to proteins in any databases. The majority of genes families with homology to proteins in databases are homologous to genes known from unicellular taxa and were not limited to multicellular taxa. Using two methods for identifying altered selection over 200 genes were identified with evidence of positive or accelerating selection.

Amongst these genes were multiple genes related to the ubiquitin proteasomal pathway (UPP) - including all steps in the pathway. These results suggest a correlation with the origin of multicellularity and because this pathway is known to be critical for regulating the cell cycle, the association with the origin of multicellularity is most likely related to the cell cycle regulation. Regulation of key cell cycle proteins through ubiquitin targeting and subsequent degradation may be central to the evolution of a genetic programme for controlling the number of cell divisions in the multicellular volvocines. Furthermore, other genes involved in the cell cycle such as cyclins and retinoblastoma gene were explored in T. socialis and these findings are described. Additional gene families with known roles in volvocine developmental traits are described including genes such as the *regA*-related genes, pherophorins and matrix metalloproteases. Fasciclin-domain containing proteins from the lineage were explored and homologues of fasciclin genes with roles in cell adhesion and embryo inversion in Volvox carteri were not identified in the genomes of other taxa. The manuscript by Arakaki et al (2017) describes the molecular mechanisms underlying the evolution of incomplete cytokinesis through examining the localisation of dynamin-related protein 1 (DRP1) during cell division. Differences in DRP1 localisation to division planes in unicellular *Chlamydomonas reinhardtii* and the multicellular *T. socialis* and *G. pectorale* were observed, suggesting that altered localisation of DRP1 was important for the evolution of incomplete cytokinesis and multicellularity.

For this manuscript the nuclear genome sequence of *T. socialis* was utilised to identify the *DRP1* gene prior to the genome being publically available.

1. Yoko Arakaki, Takayuki Fujiwara, Hiroko Kawai-Toyooka, Jonathan Featherston, Pierre Durand, Shin-ya Miyagishima, Hosayoshi Nozaki. (2017). Evolution of cytokinesis-related protein localization during the emergence of multicellularity. BMC Evolutionary Biology. 17:243

2. Jonathan Featherston, Yoko Arakaki, Erik Hanschen, Patrick Ferris, Rick Michod, Bradley Olson, Hisayoshi Nozaki, Pierre Durand. The 4-celled *Tetrabaena socialis* nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocines lineage. (2018). *Molecular Biology and Evolution*. 35(4):855-870.

4 THE *TETRABAENA SOCIALIS* MATING-LOCUS

4.1 Abstract

The volvocine lineage is an ideal lineage for exploring the genetics of sexual differentiation and the evolution of gametic sexual dimorphism. The lineage spans from isogamous to oogamous gametic forms. Homothallic and heterothallic taxa are present within the lineage. Tthe molecular genetics underpinning gender determination as well as the evolution of oogamy from isogamy require a greater understanding. Volvocine mating type is determined by a multi-gene mating locus and currently the mating loci (for both mating-types) have been sequenced for three heterothallic taxa. Furthermore, the nuclear genome sequence of the homothallic *Tetrabaena socialis* is available. A major question, which may aid in the general understanding of gender determination, is what the genetic differences between homothallic and heterothallic taxa are. In this chapter, the nuclear genome sequence of *T. socialis* is explored to gain insight into the genetic differences between homothallic and heterothallic volvocine. This chapter is to be prepared as a manuscript for submission.

4.2 Introduction

4.2.1 Volvocine Sexual Reproduction

The volvocine lineage of chlorophytes is an exceptional model lineage for understanding the evolutionary transition to multicellularity (Starr 1968; Kirk and Harper 1986; Kirk 1999; Nozaki, et al. 2000; Michod,

et al. 2003; Kirk 2005; Herron and Michod 2008; Coleman 2012). The lineage is exceptional as a model because many extant taxa are known that span from unicellular to colonial multicellular forms to fully multicellular with germ/soma differentiation (Larson, et al. 1992; Coleman 1999, 2012). The lineage is also of major interest for understanding the evolution of sexual dimorphism (Kirk 2006; Hamaji, et al. 2008; Ferris, et al. 2010; Hamaji, Ferris, et al. 2013) because there are 3 major mating systems in the lineage: isogamy, anisogamy (external and internal fertilisation) and oogamy (Nozaki, et al. 2014). Furthermore, there are taxa within the lineage in which sexual reproduction requires different strains (outcrossing or heterothallism) and taxa where both mating types are produced from the same strain and mating occurs within the same strain (selfing or homothallism). The lineage is therefore exceptional for understanding various aspects of sexual evolution, however, there remain many fundamental questions to be answered. These include: (i) the molecular underpinnings that programme different genders, (ii) the molecular genetics that define the modified programs of embryogenesis between the sexual and asexual life cycles, (iii) the genetic underpinnings that define the transition from isogamy to oogamy and (iv) the molecular differences between heterothallic and homothallic taxa are not well understood. Isogamous volvocine taxa include *Chlamydomonas reinhardtii*, *Tetrabaena socialis* and *Gonium pectorale*, anisogamous taxa include *Colemanosphaera* (external fertilisation), *Eudorina* spp and *Pleodorina starrii* (both internal fertilisation) and oogamous taxa include *Volvox* spp

(Coleman 2012; Nozaki, et al. 2014). Isogamous taxa reproduce via external gametic conjugation. In the *Colemanosphaera*, which produce anisogamous gametes, male and female gametes conjugate external to the female colony whereas for other anisogamous and oogamous volvocines sperm enters the female colony and eggs are fertilised internally (Nozaki, et al. 2014). Parsimony suggests that isogamy was the program present in the ancestor of the multicellular volvocines, while oogamy has most likely evolved twice independently in two clades of *Volvox* (Nozaki, et al. 2014).

In ideal conditions such as in nutrient rich environments reproduction occurs vegetatively (asexually). In *Chlamydomonas* N-deprivation induces the sexual life cell whereas in *Volvox* the sexual life cycle is induced by species-specific and highly potent sex pheromones (extremely low concentrations can initiate the sexual life-cycle) (Sumper, et al. 1993; Amon, et al. 1998; Rodriguez, et al. 1999). In isogamous taxa sexual reproduction occurs through the fusion of opposite gametes using specialised mating organelles. Upon fusion the developing zygote is diploid (hemizygous), undergoes meiosis whereupon differentiation into an embryo ensues. Sexual reproduction in *Volvox* occurs through impregnation of a sexual female by spermatozoa that enter the large female (~2000 celled) through fertilisation pores until the sperm reach an oogonium and fuse (Cave and Pocock 1951). *Volvox* sex pheromones are diffusible glycoproteins, similar to extracellular matrix (ECM) components

(Ender, et al. 1999; Hallmann and Kirk 2000; Hallmann, et al. 2001; Hallmann 2003). Response to sex-inducers is considered to initiate a signalling cascade through the ECM that triggers the formation of gametes (Ertl, et al. 1992; Godl, et al. 1995; Hallmann, et al. 1998; Ender, et al. 1999). The sexual life cycle of *Volvox* spp differs substantially from that of its asexual/vegetative life cycle and it is one of the most derived modes of reproduction in the lineage. This is particularly the case for the development of male *Volvox carteri*, which mature into 64 to 128 packets divided roughly equally between germ and somatic cells. Females more closely resemble asexual colonies (~2000 cells and ~16 gonidia) except that female gametes are immotile and possess a higher number of gonidia than asexual colonies (~32-64) (Umen 2011). In addition to 3 varieties of gamete-types, there is also variation in the modality of reproduction. For example, amongst isogamous taxa there exists variation in the formation of mating tubules and the mode of conjugation (Nozaki 1986). In *C. reinhardtii* and *T. socialis* one gamete produces a tubule while in *G. pectorale* both gametes produce tubules that meet and join during conjugation (Nozaki 1986).

Amongst anisogamous taxa there is also internal versus external fertilisation (Nozaki, et al. 2014)..

4.2.1.1 Volvocine Mating loci

There have been numerous investigations exploring volvocine sexual reproduction and its evolution including genome level

investigations. A multi-gene region of low recombination known as the mating-locus (*MT*-locus) determines the gender of heterothallic volvocines (Ferris and Goodenough 1994; Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). Although the volvocines are haploid, due to the restrained recombination within the locus it acts as a form of sex chromosome. Three of the previously published volvocine genomes (*C. reinhardtii MT*+, *G. pectorale MT*- and *V. carteri* female) are heterothallic whereas the 4th genome - *T. socialis* (the subject of this thesis) - is homothallic and its mating locus is the subject of this chapter. For heterothallic volvocines *MT*-loci of the strain opposite to the whole genome sequenced strain were characterised separately by sequencing overlapping bacterial artificial chromosomes (BAC) clones identified by chromosome walking (Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). In all three species, the *MT*-locus shared (conserved) genes are syntenic with genes that are found in or near the *MT*-loci of *C. reinhardtii*, located on chromosome 6. For all three species examined the *MT*-loci are non co-linear between haplotypes, repeat-rich and range in size from 200-300kb in *C. reinhardtii* to 360-500kb (Ferris, et al. 2002) in *G. pectorale* (Hamaji, et al. 2016) and 1.13 to 1.15mb in *V. carteri* (Ferris, et al. 2010). The majority of genes within the *MT*-loci of *C. reinhardtii*, *G. pectorale* and *V. carteri* are shared between respective haplotypes, while others have been identified that are gender limited. At least two *C. reinhardtii* autosomal genes have migrated into the *MT*–loci of *G. pectorale* and *V. carteri* – *HSP70B* and *MAT3*/*RB* (Ferris, et al. 2010; Hamaji, et al. 2016).

4.2.1.2 Gender Associated and Sex determining Genes.

The gene with the strongest association with gender determination is the male or *minus* haplotype *minus dominance* gene (*MID*), which encodes a putative RWP-RK transcription factor and is found exclusively in male or *minus* haplotypes (Ferris and Goodenough 1997; Nozaki, et al. 2006; Hamaji, et al. 2008; Hamaji, Ferris, et al. 2013). For both *C. reinhardtii* and *V. carteri* transformation with *MID* is sufficient to reverse the mating type from *plus*/female to *minus*/male (Ferris and Goodenough 1997; Geng, et al. 2014). Pseudo-male *V. carteri* produced by *MID* transformation produce sperm packets capable of fertilising wild-type females, however, pseudo-sperm are not without defects (Geng, et al. 2014). Transformation of *C. reinhardtii MID* into female *V. carteri* does not elicit the formation of male gametes and *MID* sequence evolution appears to have resulted in an altered program (Geng, et al. 2014).

Other gender-limited genes include *MTD1* (*MT*-) and *FUS1* (*MT*+) and there are numerous species-specific gender-limited genes known from the 3 taxa sequenced to date e.g. the *C. reinhardtii MT plus*-limited *MTA1* gene and the *V. carteri* MTF limited *HMG1* and *FSI1* genes (Ferris, et al. 2010; Umen 2011). The *MTD1* and *FUS1* genes are currently only known from isogamous taxa and they are not present in intact form in *V. carteri*. The *MTD1* gene is required for repression of the *plus* program of differentiation but is less critical for gender determination than *MID* (Ferris and Goodenough 1997; Lin and Goodenough 2007). The *FUS1* gene

encodes a transmembrane protein that is important for *plus* gametes in isogamous taxa because during conjugation it becomes localised to the *plus*-type actin-laden mating tubule and where it is most likely involved in gametic recognition and fusion (Ferris, et al. 1996; Ferris, et al. 1997). The *C. reinhardtii* and *G. pectorale* nuclear genomes are ~64% GC, however, their sex-limited *MID* and *FUS1* genes have an atypically low 50% GC content in both coding regions and introns, as well as shorter than average introns (Ferris, et al. 1996; Ferris and Goodenough 1997; Hamaji, et al. 2016). A suggested mechanism for the low %GC in these genes is lack of GC-biased gene conversion (Hamaji, et al. 2016). Heteroduplex sites formed during DNA recombination or gene conversion are replaced more frequently by G or C. *MID* and *FUS1* are only present on one chromosome in the hemizygous stage of heterothallic species, so have no allele to recombine with during meiosis, reducing opportunity for GC biased conversion.

In terms of GC content, the *G. pectorale MTD1* gene resembles *MID* and *FUS1* while *C. reinhardtii MTD1* is somewhat intermediate. As a key determinant for entry into mitosis the *MAT3*/*RB* is of interest in the volvocine lineage in relation to the evolution of multicellularity. It is also of interest for the evolution of sexual dimorphism because the gene has migrated from an autosomal position in *C. reinhardtii* into the *G. pectorale* and *V. carteri MT*-loci, the modified program of gametogenesis involves an altered cell cycle program and, furthermore, gender-specific alternative

splice variants have been identified in *V. carteri* (Ferris, et al. 2010). However, if alternative *MAT3*/*RB* splice variants are important for gender determination in *V. carteri* this does not appear to be the general case for volvocine sexual dimorphism because oogamy and anisogamy predate the origin of *MAT3*/*RB* splice variants (Hiraide, et al. 2013). An autosomal gene with a role in volvocine mating is the generative cell specific 1/hapless2 (*GCS1*/*HAP2*) gene (Kawai-Toyooka, et al. 2014). GCS1/HAP2 localises to mating tubules where it mediates membrane fusion during mating (Kawai-Toyooka, et al. 2014). The *GCS1*/*HAP2* gene is an ancient eukaryotic gene, which in many lineages including the volvocines is involved in membrane fusion (Steele and Dana 2009; Kawai-Toyooka, et al. 2014). Recent evidence suggests that *GCS1*/*HAP2* expression is stimulated by gametogenesis in both gamete type whereupon it accumulates in *minus Gonium* gametes at the site of fusion but is rapidly degraded post-translation in *plus* gametes (Kawai-Toyooka, et al. 2014). The mechanism underlying the difference between post-translational processing of *GCS1*/*HAP2* is not known.

4.2.1.3 Homothallism and Heterothallism

Separate strains exist for each gender in the majority of volvocines (heterothallic) and this is the most probable ancestral character state, however, homothallic taxa (where a single strain/clone produces both gametes types) are distributed throughout the lineage and are found in all 3 of the currently defined families (Coleman 2012). For example: *T.*

socialis (from the Tetrabaenaceae) is homothallic, in the Goniaceae homothallic strains of *G. pectorale* and *Gonium multicoccum* have been characterised and in the Volvocaceae homothallic strains or species of *Pandorina*, *Eudorina*, *Pleodorina* and *Volvox* have been characterised (Coleman 2012). Differences between the genetic programs of heterothallic and homothallic taxa are not known. While the genetic underpinnings of volvocine gender determination and the evolution of sexual dimorphism are the subject of historical and active investigation the evolution of homothallism has not received as much attention. A proposed molecular mechanism for the evolution of homothallism would be the failure to form haploid offspring after meiosis such that progeny are diploid containing both *plus* (female) and *minus* (male) gametologues (Yamamoto, et al. 2017). Based on estimates of genome size this was not found to be the case for homothallic strains of *Volvox* (Yamamoto, et al. 2017). However, genome size estimation alone is unlikely to be sensitive enough to rule out a diploid "sex chromosome" such as the *C. reinhardtii* chromosome 6.

A more thorough analysis of *MT*-locus gene copy number is required to rule out chromosome-level aneuploidy. Recently a mechanism for how homothallism in *Volvox* might evolve has been proposed. Knockdown of *V. carteri MID* using RNAi was found to result in a mixed population of male and female gametes, which was speculated to be similar to how homothallism might naturally evolve i.e. *MID* expression is

insufficient to induce the male/*minus* program in naturally occurring homothallic taxa (Geng, et al. 2014). This was likened unto the mutant *MT-minus* strain of *C. reinhardtii* known as *ISO1* (Campbell, et al. 1995), which produces a mixture of *plus* and *minus* gametes that are capable of iso-agglutination but cannot self-fertilise - most probably because as a *minus* strain it lacked the *FUS1* gene (Geng, et al. 2014). Sexual induction in male *V. carteri*, which constitutively expresses *MID*, alters MID cellular localisation to accumulate within the nuclei of developing androgonidia, however, factors transporting MID to the nucleus are unknown and the cellular localisation of MID upon sexual induction is unknown in other taxa (Geng, et al. 2014). Alternative localisation of MID may be a further potential route for the evolution of homothallism in *Volvox*. Additional recent support for gene expression regulation of *MID* as the determinant of gender in homothallic species of *Volvox* comes from the analysis of *MID* expression in a homothallic strain of *Volvox africanus* (Yamamoto, et al. 2017). Spheroid-forming female colonies were found to have a reduced expression level relative to vegetative spheroids (Yamamoto, et al. 2017). It is unknown if this is causative for oogonidia formation, however, the dominant role of *MID* throughout the lineage suggests that it is likely to be important for the evolution of homothallism. In this chapter, the lineage-wide mating-locus genes and their organisation within the nuclear genome of the homothallic *Tetrabaena socialis* are explored and findings interpreted in the context of the evolution of homothallism specifically for *T. socialis* and more generally for the lineage as a whole.

4.3 Methods

4.3.1 Identification of Mating-Locus Genes and Scaffolds

The published genome and computationally predicted proteome of *Tetrabaena socialis* (Chapter 3) were used to explore the *T. socialis MT-* locus. Mating-locus genes from both genders of *C. reinhardtii*, *G. pectorale* and *V. carteri* were searched against the *T. socialis* genome using TBLASTN (Altschul, et al. 1997) as well as against computationally derived proteins using BLASTP. OrthoMCL families and reciprocal orthologues that were generated previously from the proteomes of *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* were used to aid in the identification of putative *T. socialis* orthologues (Chapter 3).

4.3.2 Synteny

For two of the larger *T. socialis* scaffolds (scaffolds 178 and 222) MAUVE (Darling, et al. 2004) was used to evaluate conserved synteny with the mating-loci (both gametologues) of *C. reinhardtii*, *G. pectorale* and *V. carteri*.

4.3.3 Polymerase Chain Reaction Confirmation of *MID*, *FUS1* and *MTD1* Gene Structures

To confirm selected exon/intron boundaries in *MID*, *FUS1* and *MTD1* as well as to confirm their presence in the *T. socialis* genome, primers (Table 4.1) were designed to amplify selected region with the

polymerase chain reaction (PCR). The anticipated sizes of amplicons were compared to the actual size identified after separation on an agarose gel. The PCR reactions were performed at the University of Arizona, Tucson, Arizona, USA. DNA from *T. socialis* NIES-571 was extracted using CTAB for purification and Phusion *Taq* polymerase was used for amplification.

Table 4.1: Primer sequences used to confirm the gene presence and structure for *MID*, *MTD1* and *FUS1* genes.

Primer	Sequence (5' to 3')
MTD1 Forward	GTGATGTTCGCCATGGACTAT
MTD1 Reverse	GAGAACCTGTCCTGCAACG
MTD1_2 Forward	TGGCTGGAATCCTGGACTACAG
MTD1_2 Reverse	AGCAACATGATGGACACCC
FUS1 Forward	TGTAACTTGGAGCCGCATAG
FUS1 Reverse 1	CTATATGCTACCACGAGACAGTTC
FUS1 Reverse 2	ATAACCACAACGTCACCTTCAG
FUS1 Reverse 5	TTGTAGGAGAGGCTTGTTTCTG
FUS1 Reverse 4	CTACCTCGTACATCCAACCTTC
FUS1 Reverse 5	AAGTCAGTGGCAGTCCATTC
MID Forward	GCTGTACGATACTCGGCATAC
MID Reverse	GTATGGTTGTGACAGGGTTCA

4.3.4 Phylogenetic Analyses

Currently, *MTD1* and *FUS1* protein sequences are only publically available for *C. reinhardtii* and *G. pectorale*. Therefore, phylogenetic analyses were restricted to MID proteins, which were available from additional taxa including from *Chlamydomonas globosa*, *C. reinhardtii*, *T. socialis*, *G. pectorale*, *Gonium viridistellatum*, *Gonium quadratum*, *Gonium multicoccum*, *Gonium octonarium*, *Pleodorina starrii*, *Volvox africanus*,

Volvox ferrisii, *Volvox reticuliferus* and *V. carteri* (Ferris and Goodenough 1997; Hamaji, et al. 2008; Hamaji, Ferris, et al. 2013; Yamamoto, et al. 2017). Alignment of MID proteins was performed using MAFFT (Katoh, et al. 2002) and phylogenetic analysis was performed with RaXML (Stamatakis 2006) (1000 bootstrap replications) using the JTT model amino acid substitution model with a gamma distribution, which were selected after model evaluation with ProtTest (Abascal, et al. 2005). An alignment of the GCS1/HAP2 protein from *Coccomyxa subellipsoidea*, *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* was performed using MAFFT and the signal peptide sequence in *T. socialis* was identified using SignalP (Petersen, et al. 2011).

4.4 Results

4.4.1 Scaffolds Housing Mating Locus Genes

The *MT* locus of the homothallic *T. socialis* was not assembled into a single scaffold. The repeat-rich nature of volvocine mating-loci renders the locus challenging to assemble using short-read shotgun sequencing. However, two scaffolds, scaffold 178 (190kb) and scaffold 222 (163kb) were identified that each house several *MT*-locus genes. Synteny with the mating-loci of both genders of *C. reinhardtii*, *G. pectorale* and *V. carteri* was evaluated for these two *T. socialis* scaffolds (Figure 4.1). No evidence for whole-duplication of the *MT*-locus was identified. Housed within scaffold 178 were 10 genes homologous to genes from either the T (telomere proximal) and R (rearranged) domains of *C. reinhardtii* (Ferris,

et al. 2002; Ferris, et al. 2010). Scaffold 178 shows synteny to both the T domain and R domain of *C. reinhardtii*, however, in *T. socialis* orthologues of *C. reinhardtii* genes from the T and R domain are interspersed. For example, the order of T domain genes *182390*, *PSF2*, *ALB3*, *182392* and *RFC4* are preserved while R domain genes *LPS1*, *DLA3*, *533872*, and *TOC34* are interspersed with T-domain genes. The hypothetical genes *MTF0716* (*V. carteri* and *G. pectorale*) and *GMTp14* (*G. pectorale*) were also found on scaffold 178.

Of interest is the presence of the gamete-specific *C. reinhardtii minus* gene *294687*, which was found on scaffold 178 of the homothallic *T. socialis*. Housed on scaffold 222 are the genes *FUM1*, *MTD1*, *SPL2*, *NMDA1* and *DRG1* as well as 2 homologues of *G. pectorale minus* "hypothetical genes" (located between *MAT3* and *WDR57* in *MT- G. pectorale*) and the *V. carteri MTM0001* gene.

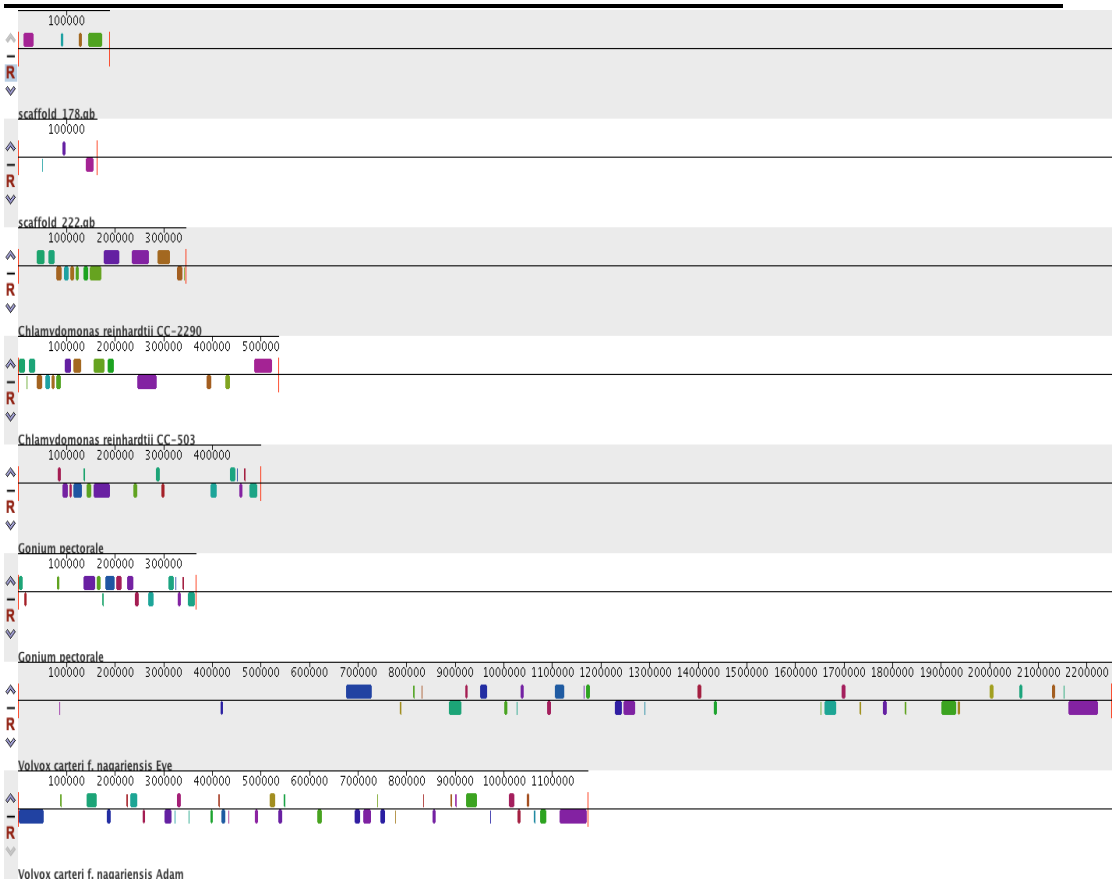


Figure 4.1: Mating locus synteny. *Tetrabaena socialis* scaffolds 178 and 222 were used to compare synteny to other volvocines. Regions of synteny conservation are identifiable. *Tetrabaena socialis* scaffold 178 shows synteny with both the T and R domains of *Chlamydomonas reinhardtii*.

4.4.2 Putative Mating Locus Genes

Gene locations were used to indicate the likelihood of a gene being housed within a hypothetical *T. socialis* mating-locus. Although a complete *MT*-locus is required for a more thorough assessment, genes found on scaffolds containing *T. socialis MT*-genes were examined to assess

whether *MT* genes were located on otherwise autosomal scaffolds. Genes located on smaller scaffolds (<100kb) were classified as unknown (Table 4.2). Using this approach, it was determined that the genes *ATPvC1*, *SeLEF*, *MME6*, *HSP70B*, *FTT2* and *AMPKR1*, which are only known from the *MT*-loci of *G. pectorale* and *V. carteri* (not *C. reinhardtii*), are located on otherwise autosomal *T. socialis* scaffolds (>120kb) and are not considered to reside within a *MT*-locus. It could not be determined from the *T. socialis* genome and *MT-locus* if genes unique to *G. pectorale* and *V. carteri MT-loci* that are not present in the *C. reinhardtii MT*-loci (*MAT3*, *WDR57*, *RPL37*, *METM1* and *ATPvL1*) are present in the *T. socialis MT*-locus. The *MAT3* gene was identified on scaffold 879. At 31kb in length it is uncertain if the *MAT3* gene resides within the *T. socialis MT*-locus, however, other genes located on scaffold 879 are not homologous to known volvocine *MT*-locus genes. The *T. socialis MT*-locus contains a mosaic of genes that are variously shared within *MT*-loci throughout the lineage. Genes shared by *C. reinhardtii* and *V. carteri* but absent from *G. pectorale MT*-loci were identified within the *T. socialis* locus (e.g. *ALB3* and *SPL2*). Six genes found in the *MT*-loci of *G. pectorale* and *V. carteri* but not in the *C. reinhardtii* mating-loci are located on *T. socialis* autosomal scaffolds.

The genes *LPS1*, *DLA3*, and *NMDA1*, which are housed within the R-domain of *C. reinhardtii* but not in the R-domains of *G. pectorale* or *V. carteri*, are found on the *T. socialis* scaffolds 178 and 222. From this

preliminary investigation into the *T. socialis MT*-locus, the locus shows more similarity to the *C. reinhardtii* haplotypes than those of *G. pectorale* or *V. carteri*.

Table 4.2: Genes likely or unlikely to reside within a hypothetical *Tetrabaena socialis* mating-locus based on gene location. See text for a detailed description of how genes were classified.

Gene Symbol	*Tetrabaena* Scaffold/s	Likely *MT* Loci?	Present in Taxa
155027	No hit		*Chlamydomonas*
182389	No hit		*Chlamydomonas*
182390	178	Yes	*Chlamydomonas*
294687	178	Yes	*Chlamydomonas*
522872	178	Yes	*Chlamydomonas*
522875	450	Unknown	*Chlamydomonas*
CGL70	48	No	*Chlamydomonas*
DLA3	178	Yes	*Chlamydomonas*
GCSH	23	No	*Chlamydomonas*/*Volvox*/*Gonium*
HDH1	450	Unknown	*Chlamydomonas*
HRGP1	41	No	*Chlamydomonas*/*Volvox*
LEU1S	208	No	*Chlamydomonas*/*Volvox*/*Gonium*
LPS1	178	Yes	*Chlamydomonas*
MADS2	4355	Unknown	*Chlamydomonas*/*Volvox*
NMDA1	222	Yes	*Chlamydomonas*
PDK1	267	No	*Chlamydomonas*
PKY1	2895	Unknown	*Chlamydomonas*/*Volvox*
PR46a	No hit		*Chlamydomonas*/*Volvox*
PR46b	26	No	*Chlamydomonas*/*Volvox*
SPL2	222	Yes	*Chlamydomonas*/*Volvox*
SPP1C	No hit		*Chlamydomonas*
TOC34	178	Yes	*Chlamydomonas*/*Volvox*/*Gonium*
UBCH1	5087	Unknown	*Chlamydomonas*/*Volvox*
UTP1	414	Unknown	*Chlamydomonas*/*Volvox*
MID	1964	Unknown	*Chlamydomonas*/*Volvox*/*Gonium*
MAT3	879	Unknown	*Volvox*/*Gonium*
HSP70B	7	No	*Volvox*/*Gonium*
ATPvC	312	No	*Volvox*/*Gonium*
EFG8	295	No	*Volvox*/*Gonium*
FTT2	503	Unknown	*Volvox*/*Gonium*

SeLEF	127	No	*Volvox*/*Gonium*
MME6	135	No	*Volvox*/*Gonium*
AMPKR1	247	No	*Volvox*/*Gonium*
METM	1465	Unknown	*Volvox*/*Gonium*
PTC1	1473	Unknown	*Volvox*/*Gonium*
RPL37A	1021	Unknown	*Volvox*/*Gonium*
SPS1	371	Unknown	*Volvox*/*Gonium*
WDR57	1019	Yes	*Volvox*/*Gonium*
PSF2	1019	Yes	*Volvox*
MTF1436	371	Unknown	*Volvox*
GMTp14	178	Yes	*Gonium*
Minus_MT.g1279	222	Yes	*Gonium*
Minus_Mt.g1280	222	Yes	*Gonium*
DRG1	222	Yes	*Chlamydomonas*/*Volvox*/*Gonium*

4.4.2.3 Gender-limited Genes

The gender-limited *minus*-dominant *MID*, *FUS1* and *MTD1* genes as well as the *C. reinhardtii* minus associated *294687* gene are present in the *T. socialis* genome. The *MID*, *FUS1* and *MTD1* genes were manually curated for gene structure. Complete proteins could not be reconstructed for the aforementioned genes for regions of low sequence similarity, however, conserved regions were reconstructed. PCR results to confirm the presence of the *MID*, *FUS1* and *MTD1* genes as well as to confirm selected exon/intron boundaries were mostly in agreement with computationally predicted gene models. However, most likely due to assembly error, exons 4-7 of the *MTD1* gene were located on scaffold 2556 and not on scaffold 222 where exons 1-3 and exons 8-12 are located. PCR to connect exons 7-8 was unsuccessful. MID proteins clustered according to organismal phylogenetic relationships and MID

proteins from homothallic taxa (*V. aureus*, *V. ferrisii*, *G. multicoccum* and *T. socialis*) do not reveal any special phylogenetic relationship (Figure 4.2). It is noted, however, that when compared to other volvocine proteins that are orthologous (e.g. MAT3) the scale bar indicates that MID proteins are relatively divergent. As per *C. reinhardtii* and *G. pectorale* gender-limited genes besides *MID* identified in *V. carteri* were not present in the *T. socialis* assembly.

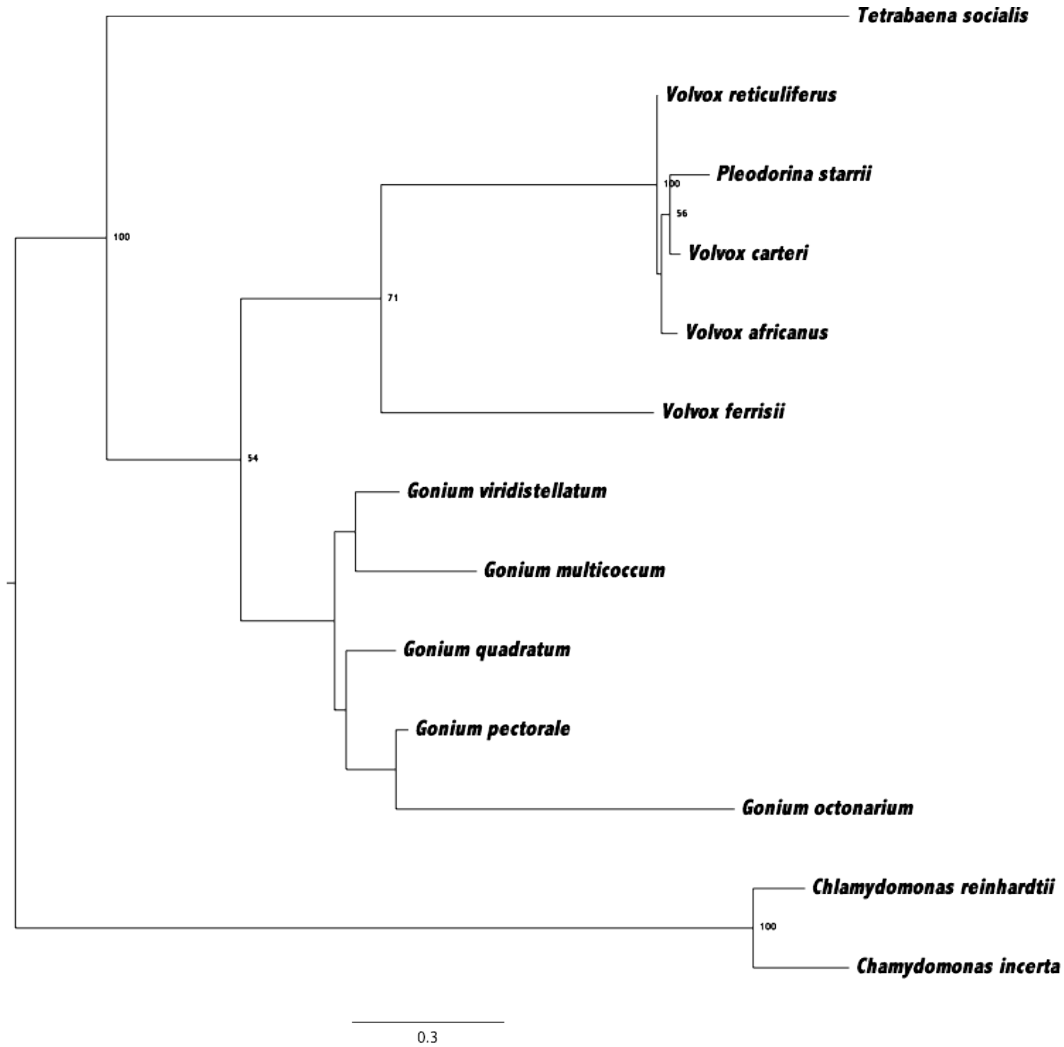


Figure 4.2: Mid-rooted maximum likelihood phylogeny of MID proteins. Bootstrap values above 50% are reported. The phylogenetic

relationship of MID proteins closely resembles that of organismal taxonomic relationships. Relative to the scale bar the distance between MID proteins is large when compared to many other volvocine proteins.

The GC content of *T. socialis MID*, *MTD1* and *FUS1* was examined (Table 4.4.2.3). As per *C. reinhardtii* and *G. pectorale* the GC contents of *FUS1* and *MID* are both lower in exons and introns than the background genomic GC content. However, the GC content of *MTD1* is similar to that of the genome GC content (~66%).

Table 4.3: GC content of *MTD1*, *FUS1* and *MID* introns and exons

	Chlamydomonas reinhardtii	***Tetrabaena socialis***	***Gonium pectoral***	***Volvox carteri***
Genome	65%	66%	65%	56%
MTD1	55%	67%	49%	NA
Introns	69%	73%	55%	
Exons				
FUS1	51%	49%	47%	NA
Introns	48%	51%	51%	
Exons				
MID	50%	51%	45%	45%
Introns	51%	51%	53%	53%
Exons				

The *HAP2*/*GSC1* gene was identified in *T. socialis* and as per *C. reinhardtii*, *G. pectorale* and *V. carteri* the gene is located in an autosomal region on *T. socialis* (scaffold 12). The *HAP2*/*GSC1* gene structure was manually curated and conserved regions were reconstructed based on

homology. The signal sequence, the HAP2/GSC1 domain and the transmembrane domain as identified by Kawai-Toyooka (2014) were identified in the *T. socialis* protein and therefore the domains required for HAP2/GSC1 function are present (Figure 4.3).

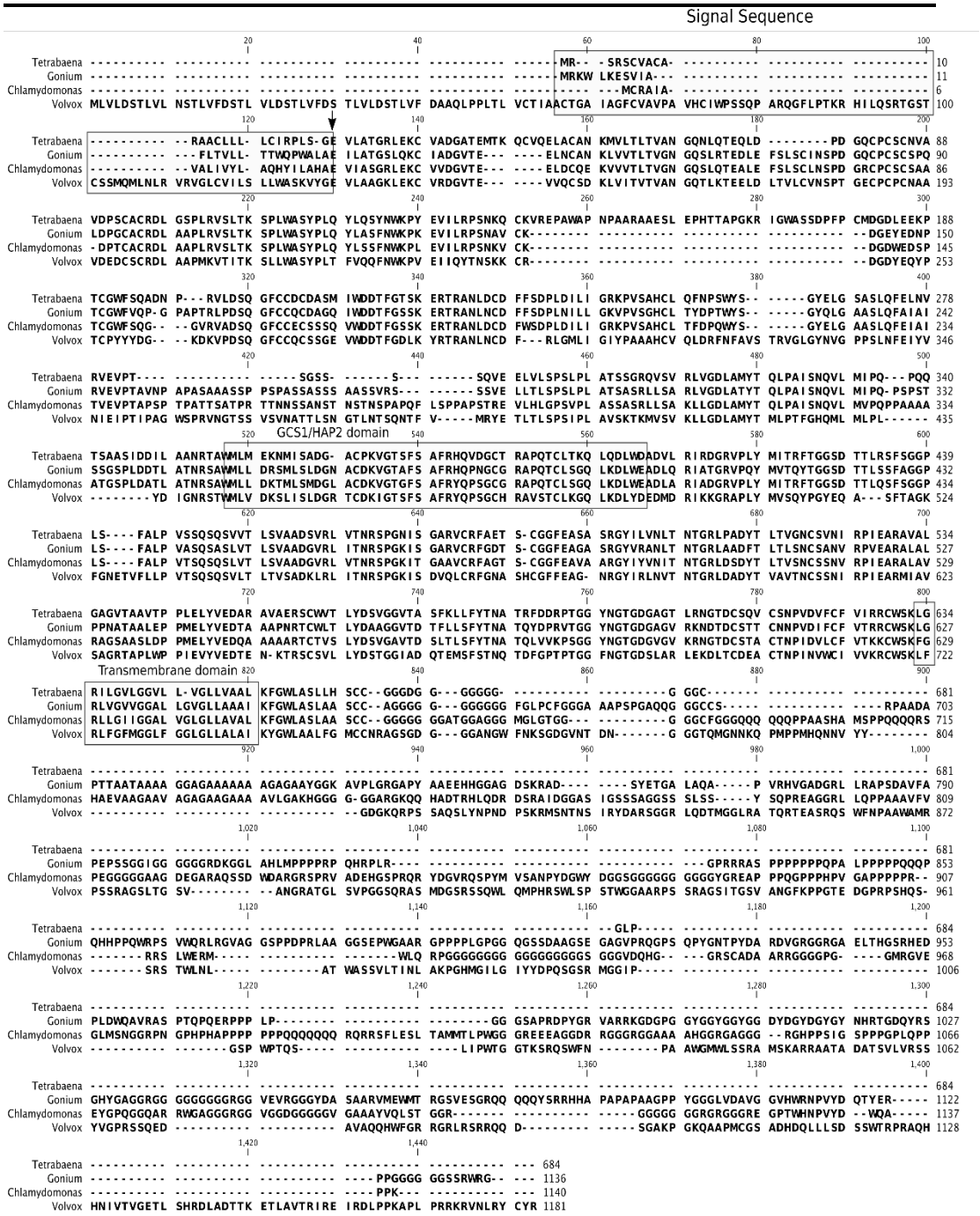


Figure 4.3: MAFFT alignment of HAP2/GSC1 proteins from *Chlamydomonas reinhardtii*, *Tetrabaena socialis*, *Gonium pectorale* and *Volvox carteri*. SignalP identified the signal peptide excision site at position 27 of the *Tetrabaena socialis* protein. The C-terminal region of the

Tetrabaena socialis protein was not reconstructed because the region is not well conserved.

4.5 Discussion

Numerous conserved *MT*-locus genes were identified within the *T. socialis* assembly. In sum, assembled scaffolds that housed a number of *MT*-locus genes were more similar to those of *C. reinhardtii* than to the *MT*-loci of *G. pectorale* and *V. carteri*. Several genes that have migrated into the *MT*-loci of *G. pectorale* and *V. carteri* do not appear to be present in the *T. socialis MT*-locus. There were no genes that are found in all three previously sequenced taxa that were not identified within the assembly, although certain of these genes in *T. socialis* reside within autosomal regions. Throughout the lineage *MT*-loci have undergone substantial syntenic rearrangements including multiple gene migrations in and out of the *MT*-locus (Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016). Many of these migrations have so far been found to be species-specific revealing a dynamic history of the locus (Ferris, et al. 2002; Ferris, et al. 2010; Hamaji, et al. 2016).

Numerous points of evidence show that although *MID* is the dominant determinant of gender throughout the lineage that gender developmental programs are complex (Ferris and Goodenough 1997; Lin and Goodenough 2007; Hamaji, et al. 2009; Ferris, et al. 2010; Umen

2011; Hiraide, et al. 2013; Geng, et al. 2014; Kawai-Toyooka, et al. 2014; Yamamoto, et al. 2017). Evidence for the complexity of regulation includes (but is not limited to): (i) the abnormal features of *V. carteri* pseudo-males produced by *MID* transformation into females, (ii) *V. carteri MID* is constitutively expressed and is not only expressed upon sexual induction (although the cellular-location of MID proteins change upon sexual induction) (Geng, et al. 2014), (iii) there are many gender limited genes housed within volvocine *MT*-loci (particularly in *V. carteri*), and (iv) when comparing gene expression between genders there are a number of genes differentially expressed (Ferris, et al. 2010). Furthermore, and somewhat paradoxically, *MTD1* knockdown in a *minus* strain of *C. reinhardtii* prevents gamete formation. However, *MTD1* is not required for a *MID*-transformed *plus* strain, which naturally lacks *MTD1*, to produce viable *minus* gametes (Ferris and Goodenough 1997; Lin and Goodenough 2007). Together, these findings are suggestive of a complex interplay between the activation of *plus* and *minus* programs that has been significantly adapted throughout the lineage where *MID* is the dominant determinant but where the genetic background and other regulatory mechanisms also play an important role.

4.5.1 GC Content

The *C. reinhardtii* and *G. pectorale* nuclear genomes are ~64% GC, however, their sex-limited *MID* and *FUS1* genes have an atypically low 50% GC content (Table 4.4.2.3) in both coding regions and introns, as well

as shorter than average introns. A suggested mechanism for the low %GC in these genes is a lack of GC-biased gene conversion (Hamaji, et al. 2016). Heteroduplex sites formed during DNA recombination or gene conversion are replaced more frequently by G or C. *MID* and *FUS1* are only present on one chromosome in the diploid (hemizygous) stage of heterothallic species, so have no allele to recombine with during meiosis, reducing opportunity for GC biased conversion. The *G. pectorale MTD1* gene resembles *MID* and *FUS1* while *C. reinhardtii MTD1* is somewhat intermediate (Hamaji, et al. 2016). The %GC of the *T. socialis* nuclear genome is 66%. *Plus* and *minus T. socialis* gametes carry the same *MT*- locus, so zygotes will contain two alleles of *MID*, *FUS1* and *MTD1,* which could recombine during meiosis. The GC-biased gene conversion explanation would predict that these three genes should have typical %GC. The *MTD1* gene, with 73% GC in exons and 67% in introns is consistent with that expectation. However, the %GC in *T. socialis MID* and *FUS1* are much the same as in *C. reinhardtii* and *G. pectorale*. There are three potential explanations: (1) GC-biased gene conversion is not the explanation for low %GC; (2) *MID* and *FUS1* still reside in a recombination cold spot even though pairing is possible or (3); *T. socialis* evolved from a heterothallic ancestor relatively recently.

4.5.2 Homothallism in the Volvocine Lineage

The presence of both *plus* and *minus MT* genes such as *MID* (*MT*- limited or male specific), *MTD1* (*MT*- limited) and *FUS1* (*MT*+ limited) in

the genome of *T. socialis* is important for understanding how this taxon is capable of self-fertilisation. The hypotheses that homothallism may have evolved in *Volvox spp* from a male strain with reduced expression of *MID*, altered regulatory expression control of *MID* or localisation of MID might be correct for *Volvox spp* or related oogamous taxa but cannot account for the evolution of homothallism in isogamous taxa where the *FUS1* gene is still required for membrane fusion during conjugation-type fertilisation (Ferris, et al. 1996). It has been noted that homothallic strains or species are less common amongst colonial volvocines than in *Volvox spp* and, furthermore, it has been suggested that some colonial species identified as homothallic may in fact be producing asexual parthenospores rather than zygotes because few instances of homothallic zygotes germinating have been observed (Coleman 2012). Potentially, zygote germination may not have been observed as frequently in homothallic colonial taxa than in members of the Volvocaceae because they have been less well studied. Zygote germination has been observed in *T. socialis* (Nozaki 1986). In *T. socialis MID*, *MTD1* and *FUS1* are present, which suggests that some form of gene shuffling between *plus* and *minus* haplotypes was involved. This was most probably via an unequal translocation event rather than an equal translocation event, which would be characterised by duplicated copies of *MT*-locus genes.

Whether *MID* and *MTD1* migrated to the *plus* haplotype or *FUS1* migrated to the *minus* haplotype cannot be determined. The most

parsimonious route would be the transfer of *FUS1* into the minus MT-locus because in addition to the presence of *MID* and *MTD1* the *C. reinhardtii* minus gene *294687* is present in *T. socialis* bringing the number of *minus*-specific genes in *T. socialis* to 3 while *plus* specific genes number only 1 (*FUS1*). Either way, the most probable cause for the evolution of homothallic reproduction in the colonial volvocines is via shuffling of genes between gametologues, which results in a haploid progeny containing all the major determinants of gender: *MID*, *MTD1* and *FUS1*.

4.5.2.1 Pseudo-Mating types

What is curious about the sexual expression of *T. socialis* is that while it is homothallic only one gamete bears a mating tubule (dioecy), which suggests that some regulatory mechanism produces pseudo *plus* and *minus* gametes (Nozaki 1986). This is the same as in *C. reinhardtii* but different from *G. pectorale* where both gametes produce mating tubules (Nozaki 1986; Mogi, et al. 2012). Besides heterothally and homothally, throughout the lineage there is variation relating to uni-directional and bi-directional mating. While *T. socialis* carries both *plus* and *minus*-specific genes it remains unknown how pseudo-gamete types are determined.

With coding potential equal between gametes post-translational or transcriptional regulatory processes are the most likely processes to explain phenotypic differences between *T. socialis* gametes and activation

of either program is presumably random. Potentially the pseudo-*minus* gamete may express *GCS1*/*HAP2* and transport the protein to the sight of membrane fusion while the *pseudo-plus* strain may express *FUS1* and degrade *GSC1*/*HAP2* production. How this is regulated is unknown but as per the knockdown of *MID* expression in *V. carteri*, differential regulation might occur through the low level of expression of a particular gene (e.g. *MID*), which may result in a variable program of gametogenesis. Whatever the mechanism, it is unrelated to gamete size as both *T. socialis* gametes are equal in size (Nozaki 1986). A better understanding may come from studying *FUS1* and *GCS1*/*HAP2* expression and localisation in *T. socialis* gametes. There remain many unanswered question regarding the molecular underpinnings of mating in *T. socialis* and it is expected that the genetic resources presented herewith may facilitate future investigations into sexual reproduction in *T. socialis*.

4.5.3 Evolutionary Implications of the Transition to Homothallic Reproduction

Homothallic reproduction in the volvocine lineage can be considered as akin to the loss of true sexual reproduction. Remarkably, despite the ubiquitous distribution of sexual reproduction amongst eukaryotes (Speijer, et al. 2015)(the last universal common ancestor of eukaryotes most likely reproduced sexually) the origin of sexual reproduction remains a mystery and more pressingly even the selective

forces that ensure the retention of sexual reproduction in most lineages is still the subject of debate (Colegrave 2012; Hartfield and Keightley 2012).

Sexual reproduction is costly. Maynard Smith coined the term “the 2-fold cost of sex” or cost of males, which describes how asexual organisms reproduce at double the rate of sexual organisms (see Colegrave (2012) for a simple explanation) (Maynard Smith 1978). Recombination can break-apart successful genotypes, is disadvantageous when heterozygosity is favoured, mate finding can be energetically costly, sexual-traits may be costly to fitness (e.g. peacock tails are not well camouflaged) and sexual competition can be highly costly (Colegrave 2012; Hartfield and Keightley 2012). A few theories have been developed to explain the advantages of sexual reproduction. These include: the Red Queen hypothesis (e.g. parasitic evasion) (Van Valen 1973), greater genetic diversity hypothesis, DNA damage repair hypothesis (Bernstein, et al. 1984, 1985b, a), breakup of linked loci and the deleterious mutational load hypothesis (Hartfield and Keightley 2012).

At present, the most widely accepted primary explanation for the retention of sexual reproduction in most instances is the increased genetic diversity that occurs through recombination (Colegrave 2012; Hartfield and Keightley 2012). This is also one of the oldest explanations for sex and relates to hypotheses such as the Red Queen hypothesis (Van Valen 1973). Reversion to homothallic mating is common in fungi and transitions

between mating-types have been well studied in many fungal lineages (Lee, et al. 2010; Metin, et al. 2010; Gioti, et al. 2012; Nieuwenhuis, et al. 2013). As per the volvocines fungi demonstrate a myriad of sexual phenotypes (Coleman 2012; Hanschen, Herron, et al. 2017). In fact the number of variations seen in the mode of sexual reproduction as well as sex determination mechanisms (Bachtrog, et al. 2014) across eukaryota is remarkable considering that sex is considered to be an ancestral trait of eukaryotes (Speijer, et al. 2015). Selfing is also found in the plant (self-fertilisation) and animal kingdoms (parthenogenesis). In some instances, self-fertilisation may become the dominant or exclusive form of reproduction in these lineages. It is not certain what the relative importance of adaptive versus non-adaptive forces are for spreading the loss of sexual reproduction through a population. The evolutionary implication for loss of heterothallic reproduction is a trade-off between the cost of sexual reproduction versus the consequent loss of diversity and other benefits of sexual reproduction.

Loss of sexual reproduction may be adaptive in the short term because organisms are able to colonise new environments from small populations without the need for finding the opposite strain i.e. from a single individual. Furthermore, another example of adaptive cause for the loss of sex might be that a successful genotype would not be disrupted by recombination.

The general assumption is that the transition to self-fertilisation/homothallic reproduction is (i) irreversible (Takebayashi and Morrell 2001; Gioti, et al. 2012) and (ii) that the loss of sexual reproduction limits further evolutionary developments and will in most instances ultimately result in extinction due to a loss of genetic variability (Takebayashi and Morrell 2001). Within the volvocine lineage there is no evidence that homothallic taxa have a higher extinction rate or a lower level of diversity and reconstruction of ancestral character states suggest that there may have been reversions from homothallic mating to heterothallic mating (Hanschen, Herron, et al. 2017). Besides the volvocine lineage there are many examples where extinction is not the outcome even over a long periods of time such as the Bdelloid rotifers, which transitioned to parthenogenic reproduction >25 million years ago and have undergone extensive subsequent diversification (Colegrave 2012). Diversification and extinction rates of asexual versus sexual organisms have not been assessed sufficiently well to conclude that transitions to asexual/homothallic reproduction result in evolutionary "dead-end lineages". In plants obligate self-fertilisation has been shown to result in higher rates of extinction but also higher rates of diversification (Wright, et al. 2013). It is most likely that neutral evolutionary processes have been responsible for the loss of sexual reproduction in certain lineages. As per adaptive responses that favour the loss of sex a low effective population size also favours spreading of traits through drift (Lynch and Conery 2003; Lynch, et al. 2016).

In the long-term the loss of sexual reproduction might promote extinction, however, in the short-term especially for facultative sexual organisms this loss may only be mildly deleterious. The neutral spread of asexual reproduction may be a good example of how a “mildly” deleterious trait can be spread through a population. In the volvocines the loss of heterothallic reproduction results in the loss of one of the major benefits of sexual reproduction, which is to increase genetic diversity, because although meiotic recombination occurs after mating genetic shuffling between individuals of the same clonal strain has a limited effect on genetic diversity. Possibly, gamete formation in homothallic volvocines is merely a vestigial trait that is initiated by environmental cues such as nitrogen deprivation. Alternatively the trait is preserved because it serves to remove damaged DNA or because sexual reproduction serves to initiate the formation of resistant zygospores (Bernstein, et al. 1984; Hoelzer and Michod 1991; Colegrave 2012; Hartfield and Keightley 2012). Homothallic reproduction evolves readily in the volvocine lineage and the reasons for this are not known. As haploid organisms it has been suggested that inbreeding depression, which is a cost to diploid taxa that undergo selfing, is ameliorated in the volvocine lineage (Hanschen, Herron, et al. 2017). Sexual reproduction in the volvocines is facultative and, therefore, the spread of homothallic reproduction may have occured through neutral evolutionary processes. At the molecular level the predisposition to lose heterothallic reproduction might be due to the strong effect size of *MID*

and its association with the *minus*/male program. However, more work is needed to clarify this question.

To date more homothallic taxa are known from amongst larger volvocines and homothallism is associated with larger organismal size (Coleman 2012; Hanschen, Herron, et al. 2017). The ultimate cause for this might be due to a relaxation of selective pressure from predation (e.g. by rotifers). The smaller taxa are more vulnerable to predation than larger taxa, which have to a greater extent escaped the Red Queen predator/prey loop that requires a large base of genetic diversity in order to innovate and survive. Alternatively, larger organismal size may be viewed as a cost instead of a benefit because larger organisms tend to have low effective population sizes (Lynch and Conery 2003) or this may be due to insufficient sampling or higher extinction rates in colonial taxa.

5 FINAL THOUGHTS AND CONCLUSIONS

5.1 General

The volvocine lineage is one of the most amenable lineages for modelling the origin of multicellularity and the evolution of organismal complexity. The fundamental questions and points of discussion are, including but not limited to: (i) theoretical considerations related to levels of selection, transitions in individuality, cooperation and conflict during transitions in individuality (e.g. (Michod, et al. 2003; Hanschen, Davison, et al. 2017)), evolutionary trade-offs (Michod, et al. 2006; Solari, Kessler, et al. 2006), the evolution of emergent properties and the evolution of altruism (Nedelcu and Michod 2006); (ii) the physics of morphological evolution (Solari, Kessler, et al. 2006; Solari, et al. 2008; Solari, Drescher, Ganguly, et al. 2011; Solari, Drescher and Goldstein 2011; Brumley, et al. 2015; Hohn, et al. 2015); (iii) the genetic and genomic underpinnings of evolutionary developments (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016); and (iv) the driving forces/selective pressures behind the evolutionary of new developments (Solari, Kessler, et al. 2006). An important question that arises when using a model system is whether or not findings from are of general significance? In the case of a model lineage such as the volvocines, can findings from this lineage be extended more generally to the evolution of multicellularity in other lineages? Findings from exploring a model lineage should be considered in a broader context.

Using genomes of extant volvocine taxa the aim of the thesis was to better understand the genetic mechanisms that gave rise to multicellularity in the volvocines. From this thesis a number of observations are made, which might apply more generally to the evolution of multicellularity and biological complexity:

- Population genetic effects and neutral evolution played a significant role in volvocine genome evolution. Population genetic data for volvocine algae are limited but available evidence suggests an inverse correlation between effective population size and organismal complexity. Volvocine population genetic data have largely come from organelle genomes since the organelle genomes are more readily sampled than whole nuclear genomes. Population data from volvocine organelle genomes suggest that various aspects of organelle genome evolution have been driven by neutral evolutionary processes, for example: multiple transitions between linear (or linear fragmented) and circular mtDNA conformation as well as inflated organelle genome size, which correlate inversely with effective population size. Organelle genomes demonstrate that neutral evolutionary processes have most likely affected various aspects of genome evolution.

Current theories that have been predominantly informed by genomic evidence suggest that neutral evolutionary processes have driven various aspects of genome and organismal evolution, such as the evolution of introns (Koonin, et al. 2013; Koonin 2016), spliceosomal

machinery (Koonin, et al. 2013; Koonin 2016), the expansion of genome size (Lynch and Conery 2003) and interactome complexity (Fernandez and Lynch 2011). Such neutral processes may be particularly applicable towards understanding major evolutionary transitions. The MTEs by definition involve substantive increases in complexity and neutral evolutionary processes may facilitate this through the fixation of traits/genes, which may not have an initial selective advantage but might play a role in future evolutionary processes. The population effects of a transition in individuality remain unknown. One possibility is that newly formed unstable transitional collectives undergo population bottlenecks thereby weakening the power of selection and strengthening the effect of neutral evolution. The effect of this would be to increase the role of chance with a potential effect on molecular evolvability. Further work in the volvocine algae will provide insights into the relative power of selection during transitions in individuality. Genetic resources such as genome sequences will facilitate these efforts.

- As has previously been reported and was found to be the case in this thesis, the evolution of multicellularity in the volvocine lineage was not associated with major increases in proteome complexity (Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). This raises the question of whether or not this is the general case for the evolution of multicellularity. In other lineages, more complex proteomes are associated with multicellular living, however, the time span separating multicellular taxa from related unicellular taxa in most lineages renders it near

impossible to determine if the increase in proteome complexity was initially associated with the evolution of multicellularity. In this regard the volvocine algae provide a model for exploring a more recent transition and findings do not indicate that the transition to simple multicellularity required much by way of protein coding innovation. It is likely that observed differences in proteome complexity between the unicellular relatives of multicellular lineages such as plants, animals (e.g. (King, et al. 2008; Moroz, et al. 2014)) and fungi are due to later diversification and lineage specific developments in the multicellular taxa.

- In this thesis, multiple genes with altered selection pressures were associated with the evolution of multicellularity. It is worth noting that the model of Maliet et al (2015) (Maliet, et al. 2015) for how simple volvocine colonies such a *Tetrabaena socialis* may have evolved suggests a gradual transition to colonial living. During a gradual transition the expectation would be that multiple genes might be under altered selection, each gene under altered selection potentially resulting in a relatively small phenotypic effect but collectively resulting in significant alterations to organismal phenotype where the resulting phenotype arises as an emergent property. Data from this thesis is in favour of multi-gene involvement. Speculatively, this is probably the case for most complex evolutionary developments.

- In the volvocine lineage cell cycle modifications are associated with the transition to simple multicellularity. In other lineages cell adhesion and the development of extracellular matrix proteins might be the first and

most important modifications leading to simple multicellularity (Niklas and Newman 2013; Suga, et al. 2013). However, reconstruction of ancestral character states suggest that modification of the cell cycle into a genetic programme for controlling cell division number was one of the first steps toward the evolution of colonial multicellularity in the volvocine lineage (Kirk 2005; Herron and Michod 2008). Furthermore, volvocines divide by palintomic multiple fission division and in unicellular relatives after cell division is complete cells remain connected for a brief period of time (Coleman 2012; Maliet, et al. 2015). In this thesis a number of genetic modifications related to the cell cycle were explored. Besides the analysis of core cell cycle genes that have been highlighted in other analyses, one of the most important observations in this thesis is that all steps in the ubiquitin proteasomal pathway show evidence of altered selection in the multicellular volvocines. This pathway is known to be integral for the regulation of the cell cycle through protein degradation thereby regulating the intracellular concentration of key cell cycle proteins – “regulation through degradation” (Ciechanover, et al. 1984; Pines 1994; Chun, et al. 1996). The genetic control of cell number is a universal trait of volvocines, however, cell number differs per species. Unicellular relatives of metazoans possess many key regulators of the cell cycle (Sebe-Pedros, et al. 2016). This may indicate that regulatory programmes were important for cell cycle modifications in other lineages.

In previous analyses and, in this thesis particularly, substantial differences in volvocine non-coding DNA have been identified. Very few highly conserved non-coding DNA elements are found in whole-genome alignments. Efforts to study non-coding regions were stymied by the degree of nucleotide divergence. While little can be made of this it is known that, for example, micro-RNAs differ between *Chlamydomonas* and *Volvox* (Prochnik, et al. 2010; Dueck, et al. 2016) and differences in transcription enhancers and promoters, which mostly reside within intergenic non-coding regions, are likely to have a bearing on volvocine evolution. Given the similarity in protein coding potential regulatory processes and regulatory elements may play a significant role in how the lineage evolved. Modifications to systems of regulatory elements have been associated with the evolution of metazoan multicellularity (Sebe-Pedros, et al. 2016).

- Conserved synteny between *Chlamydomonas* and *Volvox* (Prochnik, et al. 2010) is akin to that between Humans and Chickens. The same degree of conservation is seen in *Tetrabaena* and *Gonium* (Hanschen, et al. 2016). The implications of this are as yet unknown, however, certain authors have suggested that large-scale DNA re-arrangements including at the chromosomal level are important for driving evolutionary processes and speciation (Heng 2009; Heng, Liu, et al. 2011; Zanders, et al. 2014). The function of any gene must be considered in the context of the system in which it finds itself. This in part determines the

genotype to phenotype map. The system here may refer to the genome as a whole but may include the environment and the combination of epistatic interactions. The path from genotype to phenotype is constructed by the system and is not limited to single genes. This is particularly true for complex phenotypes. There is much evidence for associations between speciation and/or phenotypic variation due to intra-nucleus chromosomal position, chromosomal availability and aneuploidies (Heng 2009; Yang, et al. 2012; Du, et al. 2014; McGrath, et al. 2014; Tan, et al. 2015). However, a systems level understanding is required to understand how such variation contributes towards morphological evolution. This may remain challenging to achieve because the path from genotype to phenotypes may for most complex phenotypes not follow a simple additive or direct route but instead the phenotype may arise as an emergent property. These features have been investigated in other lineages in relation to the evolution of multicellularity and genome-level dynamics have been associated with the evolution of multicellularity (Sebe-Pedros, et al. 2016).

The volvocine lineage is also a useful model for investigating the evolution of sexual dimorphism or oogamy. Other avenues of investigation related to sexual reproduction in the lineage include identifying the genetic basis for gender determination and the genetic differences between heterothallic and homothallic taxa. The most parsimonious account is that the ancestor of the multicellular volvocines was heterothallic and therefore homothallic taxa in the lineage most likely lost heterothallism (Hanschen,

Herron, et al. 2017). In this thesis the genetic differences between heterothallic and homothallic taxa were explored. This was undertaken through exploring the genome sequence of *Tetrabaena socialis*, which is the first homothallic volvocine taxon to be sequenced. A complete mating-locus was not constructed, however, known mating-locus genes were identified within the genome sequence of *Tetrabaena socialis*. Genes with a strong association with gender were identified and genes that in other taxa are limited to either the *plus* or *minus* mating type were identified. These finding suggest that the loss of heterothallism in *Tetrabaena socialis* was due to a recombination event where genes from one mating type were swapped with the opposite mating-type. A recombination event where multiple mating-locus genes from both genders were combined is most likely to have been required for homothallism to evolve in isogamous taxa because a number of genes are required for mating-type programming in these taxa (Ferris, et al. 1996; Ferris and Goodenough 1997; Lin and Goodenough 2007). Possibly, in oogamous taxa the evolution of homothallism may be accounted for by regulatory modifications of the sex-determining gene *MID* (Geng, et al. 2014; Yamamoto, et al. 2017). Homothallism has not been associated with lower levels of species diversification or higher rates of extinction in the volvocine lineage (Hanschen, Herron, et al. 2017). Therefore, it is uncertain what evolutionary implications there are for a transition to homothallism nor why it has it evolved relatively frequently in the lineage. There may be a number of selective advantages for homothallism,

however, neutral evolutionary processes driving the evolution of homothallism should also be considered.

5.2 Future Recommendations

The volvocine lineage has already provided valuable insights into numerous aspects of evolutionary biology. Most importantly the lineage provides insights into major evolutionary developments that have profoundly shaped life on earth. Using genomics to explore the lineage is one means to explore the mechanisms important for the evolution of facultative colonial multicellularity, the evolution of sexual dimorphism and the evolution of obligate multicellularity. There are currently efforts underway to sequence the nuclear genomes of numerous volvocines and these efforts will doubtlessly uncover new mechanistic insights. In this regard the generation of complete or near complete genome sequences using newer long-read technologies (e.g. PacBio sequencing) and scaffolding tools (e.g. optical mapping) will facilitate the analysis of whole genomes including non-coding regions. The genome sequence of *Tetrabaena socialis* provides an insight into the simplest members of lineage and is expected to be of use for future comparative genomics analyses. The *Tetrabaena socialis* genome assembly will benefit from further efforts using long-read technology to complete the genome sequence.

Despite the value of generating genome sequences for comparative analyses in the lineage there have been numerous associations made between genes and traits, however, further investigations into the mechanism and function of these genes is required. In most cases the presence of particular genes have merely been correlated with the presence of traits and therefore remain as mere candidates for further investigation. Today, there exist many molecular technologies for manipulating genes and exploring gene function not the least being the Crispr-cas9 system for generating point mutations, gene reconstruction, gene transfer, RNA silencing and other applications. Other techniques include tracing intra-cellular localisation of proteins, identifying protein-protein interactions, identifying transcription factor binding sites (e.g. ChIP-seq), total RNA gene expression profiling, expression profiling of single cells, exploring DNA methylation and a myriad of other approaches. Comparative genomic investigations generate many questions that require functional assessment in the laboratory. Perhaps at this point prior to the availability even of additional volvocine genomes mechanistic understanding gained in the laboratory might provide the most valuable insights.

6. APPENDICES

6.1 Introduction to Appendices

Chapter 3 described details of the *Tetrabaena socialis* genome and comparative genomic analyses with other volvocine genomes. The main findings of Chapter 3 were described in Featherston et al (2018). However, details of some of the analyses performed as well as additional analyses were omitted from Featherston et al (2018). In the appendices that follow are descriptions of the methods performed, results and discussions pertaining to these additional analyses (Appendix 7A-9C). Appendix 10D is a discussion that attempts to reconcile findings from genomic analyses with Kirk's 12 step program.

7. APPENDIX A: GENOME AND TRANSCRIPTOME ASSEMBLY

7.1 Methods

7.1.1 Next Generation Sequencing

Sequencing of the *T. socialis* genome and transcriptome was performed on the Illumina HiScanSQ, HiSeq 2500 and MiSeq (v2) instruments. PE sequencing of the *T. socialis* genome performed on the HiScanSQ and HiSeq 2500 were performed with 2x100bp Illumina TruSeq SBS v3 chemistry whereas MiSeq v2 instrument reads were generated with v2 2x300bp chemistry. Sequence data quality was examined using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

7.1.2 Evaluation of *de novo* Genome Assemblers

To produce the most contiguous assembly from the available data a series of *de novo* genome assembly algorithms were evaluated (Table 7.1). Besides CLC (v6.5), *de novo* genome assemblies were memory intensive and were therefore executed using a bash interface on a Hewlett-Packard HPC server. Memory requirements limited most assemblies to be performed on a 500GB memory node. At the time when assemblies were being performed, algorithms were not installed onto the HPC for general usage by a system administrator and, therefore, all algorithms first had to be compiled. Computational time to complete assemblies differed by algorithm but typically varied between 2 and 7 days

(e.g. SPAdes or JR-assembler) of computation. Certain other computational procedures also required considerable computational time. For example, with available computational resources the MAKER2 annotation pipeline ran for over a week and required three iterations to produce final annotations (nearly 4 weeks of total computing time). Initially, default parameters were used for all assemblies and subsequently parameters that could be optimised were performed. However, if a contig N50 of 2kb could not be achieved with a particular assembler, then the assembler was not optimised. Thus, not all assemblers were tested with all available data and, for example, mate-pair (MP) scaffolding was not performed for all contig-level assemblies. Nonetheless the parameters tested for numerous assemblers were extensive. Details of parameters are not discussed at great length because the aim of the assembly process was to produce the most complete and contiguous assembly, parameter optimisation was highly technical in nature and evaluations were goal orientated.

In brief, parameters that were evaluated for numerous assemblers included the optimal assembly kmer, the effect of overlapping reads, alternative trimming strategies, choice of scaffolding algorithm and where applicable various parameters related to alignment stringency (e.g. for JR-Assembler or SGA). The input data and parameters for CLC *de novo*, SGA (Simpson and Durbin 2010), Allpaths-LG (Gnerre, et al. 2011) and SPAdes (Bankevich, et al. 2012) are described in further detail. CLC and

SGA assemblies are described in Chapter 2 while the nuclear genome assembly with Allpaths-LG and SPAdes are described in Chapter 3 (Featherston, et al. 2018). A number of assemblers were not successfully installed. For example, the celera assembler could not be run on the available file system of the HPC server (Lustre file system). Assemblers that were evaluated in greater depth are included in Table 7.1.2. In addition to testing assemblers with actual sequence data of *T. socialis*, ART (Huang, et al. 2012) was used to generate simulated PE and MP data from the *C. reinhardtii* v5 genome and assembly algorithms that performed best with real *T. socialis* sequence data were tested for performance using the simulated *C. reinhardtii* data. Allpaths-LG, SGA, DISCOVAR (Weisenfeld, et al. 2014), JR-assembler (Chu, et al. 2013) and CLC (with SSPACE scaffolding (Boetzer, et al. 2011)) were evaluated using these. The simulated datasets were informative for deciding between alternative sequencing strategies.

Table 7.1. Different assemblers tested.

Tool	Principle	Pipeline
JR assembler (Chu, et al. 2013)	3' extension by jump reads and mapping	Low complexity filtering, error correction, read merging, contig assembly, scaffolding, gap-closing
CLC	De Bruin Graph	Contig assembly, assembly error correction, scaffolding
Abyss (Simpson, et al. 2009)	De Bruin Graph	Contig assembly, scaffolding
Velvet (Zerbino and Birney 2008)	De Bruin Graph	Low-coverage filtering, contig assembly, scaffolding, mapping of reads to assembly.
SGA (Simpson and Durbin 2010)	Overlap-layout-consensus string graph	Trimming, error correction, low-complexity filtering, contig assembly, scaffolding
Allpaths-LG (Gnerre, et al. 2011)	De Bruin Graph	Low complexity filtering, error correction, read merging, contig assembly, assembly error correction, scaffolding.
SOAPde novo2 (Luo, et al. 2012)	De Bruin Graph	Error correction, contig assembly, scaffolding, gap-closing
SPADES (Bankevich, et al. 2012)	Multi-Kmer De Bruin Graph	Error correction, multi-kmer contig assembly, scaffolding, assembly error-correction.
DISCOVAR (Weisenfeld, et al. 2014)	De Bruin Graph	Error correction, read merging, contig assembly, contig assembly pruning

7.1.2.1 Scaffolders

Some assembly algorithms did not include a scaffolding step, or the scaffolding process may have required the installation of many additional dependencies (e.g. SGA), or the scaffolding step of the algorithm did not utilise MP data effectively (e.g. CLC). Several independent scaffolding

algorithms were examined - including MIP (v0.6) (Salmela, et al. 2011), SOPRA (v1.4.6) (Dayarian, et al. 2010), SLIQ (v1) (Roy, et al. 2012), SSPACE (v2) (Boetzer, et al. 2011), ScaffoldScaffolder (v1) (Bodily, et al. 2016), BESST (v1.1.2) (Sahlin, et al. 2014) and Opera (v1 and 2) (Gao, et al. 2011). The various scaffolding algorithms utilised alternative mathematical approaches for scaffolding (Hunt, et al. 2014). SSPACE was found to be the simplest algorithm to implement and had minimal memory requirements and was therefore used for the basic assessment of mate-pair data e.g. for testing mate-pair insert sizes or alternative trimming strategies. However, Opera was found to produce larger scaffolds.

7.1.2.2 Post-assembly

After scaffolding, additional techniques were used for genome improvement. These techniques can be divided between gap-filling approaches (to fill in gaps between contigs within scaffolds), error reduction and techniques to merge assemblies. Gap-filling algorithms mapped reads to scaffolds and subsequently extended the mapped reads into gaps with both MP and PE data. Gaps could either be shortened or closed completely by either reinitiating assembly from the start of a gap or simply by mapping of paired-end and mate-pair reads into the void of an assembly gap. The programs PRICE (v1), GapFiller (v1.11) and Pilon (v1.8) were used for gap filling (Boetzer and Pirovano 2012; Ruby, et al. 2013; Walker, et al. 2014). PRICE was found to be of little utility. To improve gaps in the *T. socialis* assembly both GapFiller and Pilon were

utilised. GapFiller performed only gap filling whereas Pilon performed both error correction and gap filling. In relation to error correction, Pilon examined for evidence of local miss-assemblies such as insertions or deletions, the incorrect ordering or orientation of contigs within scaffolds and base-level errors (Walker, et al. 2014). Both Pilon and GapFiller were performed iteratively. The programs Metassembler (v1) and GAM-NGS (v1) were evaluated for assembly merging efficacy (Vicedomini, et al. 2013; Wences and Schatz 2015). GAM-NGS was tested on multiple SOAPdenovo2 assemblies of different kmer sizes but merged assemblies showed only minimal improvement whereas Metassembler produced significant improvements and was used thereafter. The SPAdes assembly algorithm was also evaluated, and SPAdes can be considered as a form of meta-assembly because it combines the results of assemblies generated from multiple kmers.

7.1.2.3 Basic assembly statistics

Although some assemblers produced a basic assembly report, parameters such as the minimum contig size of the assembly were not consistent. In order to assess genome assemblies in a uniform manner the program Quast (2.3) was used (Gurevich, et al. 2013). Cumulative histogram plots of scaffold and contig length were produced by Quast (Figure A).

7.1.2.4 Coding-content of Genome Assemblies

The programs CEGMA (v2.4) (Parra, et al. 2007) and BUSCO (v1.1) (Simao, et al. 2015) were used to evaluate the coding potential of various *T. socialis* genome assemblies. CEGMA was used to identify a set of 248 highly conserved eukaryotic genes while BUSCO was used to identify a set of 429 conserved eukaryotic genes (typically over 90% are present in most eukaryotes genome assemblies as single-copies). In addition to running CEGMA and BUSCO directly on the genome of *T. socialis*, KOG (Tatusov, et al. 2003) domains (n=458) were scanned against the *T. socialis* proteome using HMMER3. A list KOG genes previously identified in *C. reinhardtii* (n=406) were aligned to the *T. socialis* proteome using BLASTP ($E=1x10^{-10}$) and to the *T. socialis* genome using TBLASTN ($E=1x10^{-10}$). *Tetrabaena socialis* proteins with homology to *C. reinhardtii* KOG genes (Parra, et al. 2007) were sorted to include only unique *T. socialis* proteins. Besides assessing genome completeness, BUSCO was also used to assess the completeness of predicted coding sequences (CDS). Furthermore, DOGMA was used to assess the quality of the final predicted proteome by assessing domain presence and conserved domain arrangements- the order and arrangement of domains found in conserved eukaryotic proteins from six taxa: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*. To contextualise CEGMA, BUSCO and DOGMA results for *T. socialis* these algorithms were also run against the genomes/proteomes of *C. reinhardtii*, *G. pectorale*, *P. starrii* (a

basic draft genome) and *V. carteri*. Furthermore, CEGMA and BUSCO were run against 2 available but fragmented draft genome assemblies. These assemblies were for a plant (*Rosa* spp) and a metazoan (*Steinernema MCB28*).

7.1.3 Genome Cleaning

The assembly produced from the final round of pilon – hereafter the final assembly - was inspected using tblastn (Altschul, et al. 1997) to identify regions of chloroplast DNA (cpDNA), mitochondrial DNA (mtDNA) and the Illumina control lambda *phiX174*. The rate of potential bacterial contamination was identified in the *T. socialis* using a kmer-based searching algorithm known as Kraken (v0.10.5) (Wood and Salzberg 2014). Kraken was unleashed to scan the *T. socialis* genome against a database of kmers (mini-kraken_20141208) containing bacterial and viral sequence kmers. The overall rate of sequence contamination was compared to that of *C. reinhardtii* (Merchant, et al. 2007), *G. pectorale* and *V. carteri* and found to be similar.

7.1.4 Transcriptome Assembly

7.1.4.1 Data Trimming

Approximately 75 million paired-reads (150 million total) were generated for the *de novo* transcriptome of *T. socialis* from a single RNA sample. Quality and adaptor-content trimming was performed using Trimmomatic (Bolger, Lohse, et al. 2014). Quality trimming was performed

to exclude leading and trailing (start and end of sequence) data that fell below a minimum quality of 5 and reads shorter than 36bp were removed (Haas, et al. 2013).

7.1.4.2 Trinity assembly

De novo transcriptome assembly using Trinity (v1- 2013_08_14) and Trinity 2 (v2.2.0) (Haas, et al. 2013) was performed. A single-stranded library type was specified with jaccard-clipping applied to minimise erroneous mergers of transcripts.

7.1.4.3 Bridger *de novo* assembly

Bridger *de novo* (Chang, et al. 2015) followed a similar structure of commands to those of Trinity. A single-stranded library type was selected and two kmers of 25 and 31 were tested.

7.1.4.4 Transcriptome evaluation

The program TransRate was used to evaluate the quality of *de novo* transcriptomes in the absence of a reference genome (Smith-Unna, et al. 2016). TransRate performed alignment of paired-end reads to a *de novo* transcriptome and used read-mapping to evaluate the quality of each contig in terms of: the rate of erroneous transcript mergers (chimearas), contig coverage, base level errors, insertions or deletions and misassemblies. In addition to quality measures that evaluated assembly errors TransRate evaluated the completeness of contigs by measuring the

extent that paired-end reads map off/beyond the ends of contigs i.e. one read maps to the contig while it's pair does not map to the contig. TransRate used the summary statistics of contig quality and contig completeness to generate an optimal assembly score. Contigs with and without errors were divided into separate output files designated as "Good Transcripts" and "Bad Transcripts". TransRate also performed a reciprocal blast analysis to the *C. reinhardtii* (v5.5.) proteome.

7.2 Results

7.2.1 Sequencing Output

Table 7.2 summarises data yield across the various platforms and sample types. Size selected mate-pair (MP) data were sequenced on the HiSeq 2500 with 2x100bp chemistry whereas gel-free mate-pair data were sequenced on an Illumina MiSeq v2 (2x300bp). Transcriptome sequencing was performed exclusively on the HiSeq 2500 using v4 chemistry of 2x125bp read length. Size-selected MP data included many small fragments and thus a substantial yield of data was acquired to compensate for the presence of small fragments. This was probably because the sonication with the Branson sonicator over-sheared the circularised DNA fragments. Beside the MP data that included many small fragments quality profiles indicated that PE and transcriptome data were of high quality. Raw data for *T. socialis* total DNA exceeded 600X the estimated genome size (110-150mb). However, significant amounts of data from mate-pair data were removed due the presence of adaptor-

adaptor dimers and small inserts. Furthermore, gel-free mate-pair libraries did not improve assemblies and were not used in the final assembly.

Table 7.2: Sequencing yields across various library types and platforms. Yields are reflected in both gigabases (GB) of sequence data and in read number. In the first column the source of nucleic acid material is indicated as either genome or transcriptome, which stem from DNA or RNA material.

Source	Platform/Chemistry	Library Type	Data Yield (GB)	No. of reads (Million)
Genome	HiSeq 2500 TruSeq V3 (2x100bp)	180bp PE	6,2GB	31M x2
Genome	HiScanSQ Truseq v3 (2x100bp) and HiSeq 2500 TruSeq V3 (2x100bp)	200-800bp PE	18,9GB	105M x2
Genome	HiSeq 2500 Rapid v1 (2x100bp)	6kb MP	30,2GB	151M x2
Genome	MiSeq v2 with v3 chemistry (2x300bp)	200-800bp PE	13.8GB	23M x2
Genome	MiSeq v2 with v3 chemistry (2x300bp)	1kb-19kb	17.3GB	29M x2
Transcriptome	HiSeq 2500 TruSeq v4 (2x125bp)	TruSeq Stranded mRNA	18,8GB	75M x2

7.2.2 Testing Assemblers

The PE library that was size selected to contain fragments between 500-800bp was found instead to contain 100-800bp inserts. Various assemblies produced using these data (at 2x100bp read-length) from a variety of assembly algorithms (Table 7.3) with nearly 200X coverage

could not produce contig (with some scaffolding from PE data) N50 values greater than ~3kb. Sequence reads as well as assemblies showed that the *T. socialis* genome was comprised of ~65% GC bases, which may have hampered genome assemblies. Besides the high GC content, the extent of repeat sequences was likely also hampering assemblies.

The addition of MP data improved assembly continuity but the available data could not be used to produce a scaffold N50 greater than ~23,4kb. It was not obvious how to precede using Illumina sequencing technology, which does not produce long reads capable of spanning over long regions of repetitive sequence. At this stage in the project methods for single molecule sequencing (e.g. PacBio sequencing) were either not available, were overly expensive, the systems were not well optimised and did not produce high-quality data (or algorithms to overcome poor sequence quality) and no such systems were available in South Africa. Therefore, to decide on the most appropriate choice for further Illumina sequencing a number of simulated sequencing datasets were generated from the *C. reinhardtii* v5 genome (65% GC). Three simulated libraries were generated: a 2x100bp overlapping data set (~180bp insert size), a 2x300bp data set (~500bp insert size) and a 6kb mate-pair dataset were simulated. Assemblies with these data were tested using CLC (with SSPACE scaffolding), SGA, JR-assembler, DISCOVARdenovo (SSPACE and OPERA1 scaffolding were tested), SOAPdenovo2 and Allpaths-LG. Neither SGA nor JR-assemblies completed correctly. These evaluations

indicated that both Allpaths-LG and DISCOVARdenovo generated contiguous contig N50 values, however, Allpaths-LG produced superior scaffold N50 values (Table 7.2.2). SPAdes was not tested with these data, however, it was later used with real *T. socialis* MiSeq PE data.

Table 7.3: Results from five assemblers tested with simulated *Chlamydomonas reinhardtii* data. JR-assembler could not assemble the genome from these data. DISCOVAR yielded superior contig level continuity, however, scaffolding with Allpaths-LG was superior.

Assembler	Contig N50	Scaffold N50	Total Assembly Size (mb)
SGA	669	NA	4,4
CLC-SSPACE	8 150	31 472	106
JR-Assembler	NA	NA	3
DISCOVAR	55 508	240 293	112
Allpaths-LG	54 108	313 520	99.6

On the basis of the simulated-read experiment, the results of various assembly “competitions” and compatibility with existing data, the strategy for further sequencing was to generate an overlap library for sequencing and to perform an Allpaths-LG assembly (Salzberg, et al. 2012; Bradnam, et al. 2013; Magoc, et al. 2013). Once the overlap library was sequenced Allpaths-LG was used to generate an assembly, which proved to be superior in terms of continuity than all other assemblers tested. To further improve upon the initial Allpaths-LG assembly additional

sequencing on an Illumina MiSeq was performed (2x300bp). The additional MiSeq data were used to evaluate DISCOVAR and SPAdes. The assemblers CLC and SGA were repeated with all PE data and with size selected MP data. The results of various assemblies are included in Table 7.4.

Table 7.4: Results from optimised assemblies of *Tetrabaena socialis*.

Assembler	Total contig/ scaffold size (mb)	No. of scaffolds	% Ambiguous bases	Contig / Scaffold N50 (kb)	CEGMA
Abyss	140/151	518 181	0,9%	0.7/1.7	NA
SOAP-denovo	111/119	170 198	8%	1/2.8	NA
SGA	97/127	55 887	24%	2.8/8.6	27%
CLC	112/131	75 685	15%	3.3/8.82	31%
JR-Assembler	97/142	37 034	32%	2.6/23.4	28%
DISCOVAR	90/110	17 848	18%	4.2/49.5	50%
SPAdes	104/121	16 621	14%	6/26.8	54%
Allpaths-LG	87/138	6 952	37%	5/144.4	50%

7.2.3 Final assembly

After performing an initial Allpaths-LG assembly, the assembly was repeated a further two times, each time improving the assembly by adding data or modifying the data input. The first improvement was to include *in silico* selected overlap reads from PE data 2x100bp data. The second improvement was made by adding MiSeq jump and fragment libraries. The third and final Allpaths-LG assembly produced a scaffold N50 of 144kb

and contig N50 of 5kb with a total assembly size of 138mb and total contig assembly size of 87mb. The Allpaths-LG assembly consisted of 26 476 contigs and 6 952 scaffolds with an average of 52 scaffolds per mb. The median size of gaps within scaffolds was 2 146bp. Fragment reads utilised by Allpaths-LG amounted to 70X coverage and jump reads amounted to 193X coverage.

Allpaths-LG calculated the median paired distance of the MiSeq jump library to be 433bp, the median PE HiSeq library to be 353bp and the size selected mate-pair data to be 7 053bp (57% of mates mapped between 4-8kb apart and 32% mapped as greater than 8kb). Mate-pair library preparation using automated Sage Bioscience PippenPrep pulse-field gel electrophoresis proved effective for the removal of smaller inserts. The effect of assembly merger with the outputs from other algorithms was tested using the initial Allpaths-LG assembly (Table 7.5). The output from the algorithm SPAdes was found to merge well with Allpaths-LG. This was partly because SPAdes produced a larger assembly than Allpath-LG, which aided with gap filling. As DISCOVARdenovo did not assemble more contig level data than Allpaths-LG it was not tested for merger with Allpaths-LG.

Table 7.5: Testing different assembly mergers with Allpaths-LG as the master assembly.

	Contig N50	Scaffold N50	No. of Scaffolds	% Ambiguous bases
Allpaths	5 036	107.5KB	8 301	33.2%
Allpaths-CLC	7 049	109.6KB	7 257	27.2%
Allpaths-SGA	6 738	108.8KB	7 322	27.7%
Allpaths-JR	6 400	108.8KB	7 425	26.5%
Allpaths-SPAdes	7 423	109.5KB	6 837	24.8%

The final optimised Allpaths-LG assembly as well as the SPAdes assembly were gap-filled and then merged. Subsequently, Pilon genome error correction was performed three times to reduce the number of assembly errors. Results of genome improvement methods are included in Table 7.6.

Broadly, the effect of gap filling, merging assemblies and Pilon correction was to reduce the total number of contigs (~18%) and scaffolds (~15%), to improve contig N50 (~48%), to marginally improve scaffold N50 (~2%), to reduce the gap content (~9%) and to add ~10,4mb of total contig sequence data. The final assembly was 66% GC, 135.9mb in length, half of the assembly size was made up of 263 scaffolds (the L50 value) and the largest scaffold was 697kb in length (Figure 7.1 and 7.2). It was found that 87,5% of paired-end reads mapped to the final *T. socialis* genome. This indicates that most reads were incorporated into the genome assembly. The inclusion of mitochondrial and chloroplast together with the

genome may improve the overall mapping rate, however, this was not tested.

Table 7.6: Iterative Improvements towards the final assembly.

Assembly	Contig N50	Scaffold N50	No. of Scaffolds	% Ambiguous bases	Assembled Contig Size (MB)
Allpaths-LG	5 013	144.4KB	6 952	36.5%	87.5
Allpaths-LG (Gap filled)	6 601	144.7KB	6 952	30.4%	96.1
SPAdes	6 007	26.8KB	16 621	14.5%	103.7
SPAdes (Gap filled)	8 810	26.9KB	16 621	11.4%	107.6
Merged	7 414	146.3KB	5 858	27.9%	97.9
Pilon	7 434	145.9KB	5 858	27.8%	97.9

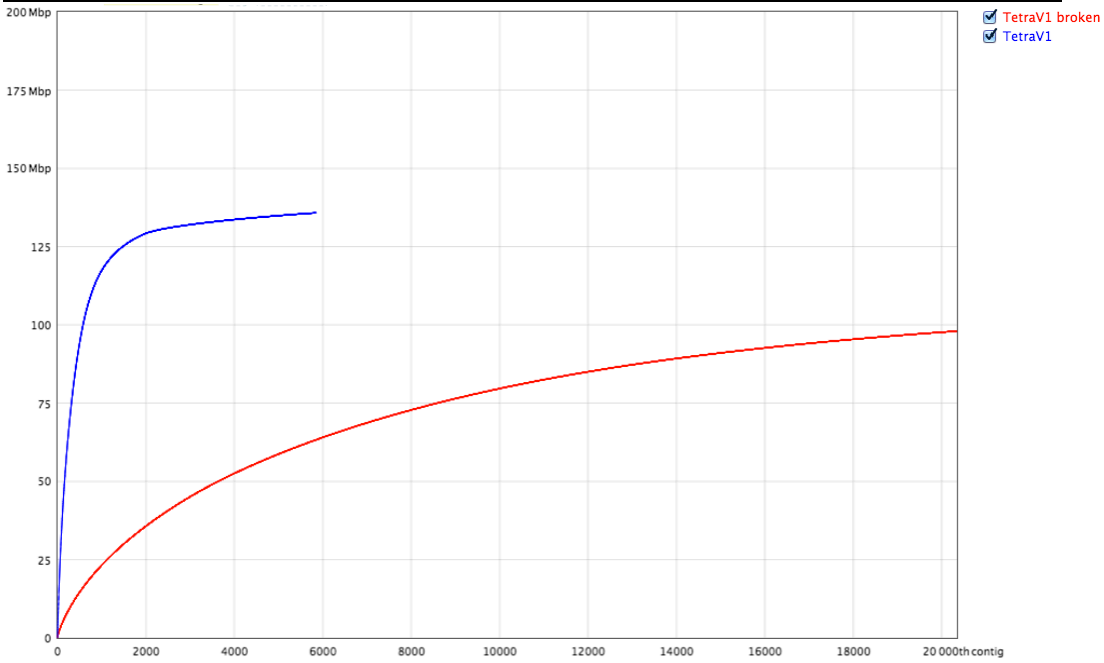


Figure 7.1: A cumulative histogram of scaffolds (blue) and contigs (red) for the final assembly of *Tetrabaena socialis*. On the y-axis is the cumulative assembly size and on x-axis is the number of scaffolds or contigs arranged from largest to smallest. Generated using Quast. The scaffold-level assembly can be seen to almost plateau once the 2000 largest scaffolds are combined.

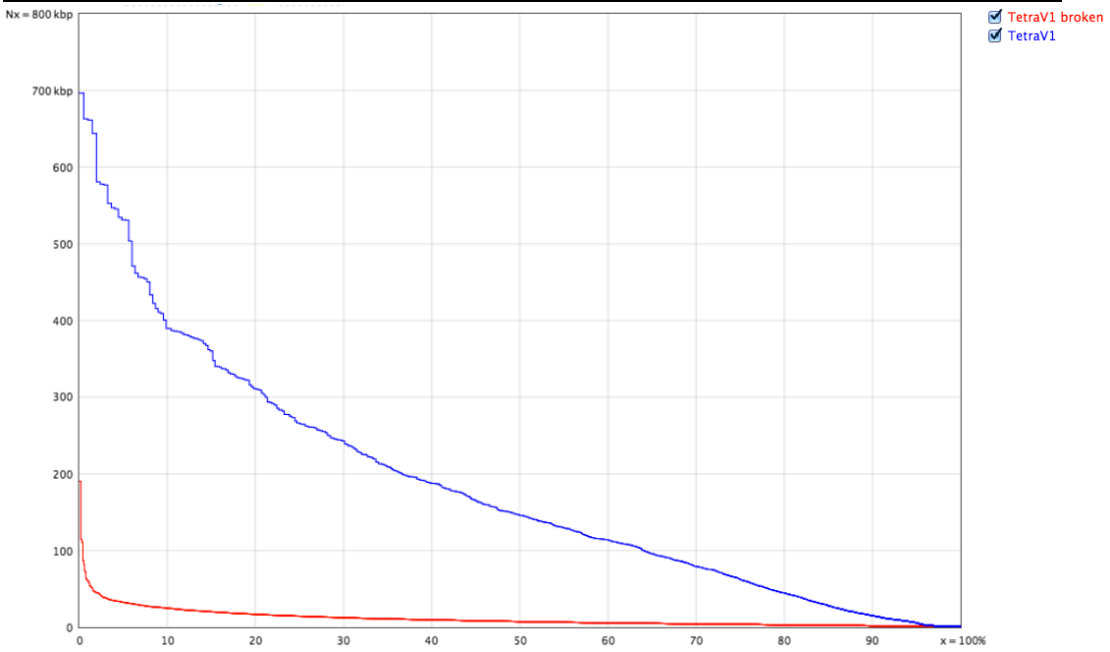


Figure 7.2: NxG diagram of scaffold and contig length (y-axis) versus percentage of genome (x-axis). As an example, 10% of the scaffold assembly is contained within scaffolds larger than 400kb. Generated using Quast.

7.2.4 Kmer Analysis

By default the Allpaths-LG pipeline performed kmer analyses to estimate the genome size, rate of sequencing bias and repeat content. In addition, kmer analyses were performed using Jellyfish and GenomeScope (Vurture, et al. 2017) and are described in Featherston et al (2018). Here, the results of Allpaths-LG kmer analyses are described in greater detail. Kmer spectrum plots were drawn from Allpaths-LG kmer analysis (for k of 25) using filtered but uncorrected fragment reads.

Kmer spectrums generated by Allpaths-LG were diagnostic of the difficulties associated with assembling the *T. socialis* genome. Kmer spectrums revealed that quality corrected reads were of high quality that did not display a high frequency of low copy number kmers (Figure 7.3). The low rate of kmer polymorphism and the absence of a second kmer peak confirm that the genome of *T. socialis* was haploid. The high GC content of the *T. socialis* genome could have resulted in substantial sequencing bias, however, kmer plots did not display this. A severe left-shift or leftwards bias in kmer frequency may be indicative of sequencing bias where certain regions of the genome (e.g. GC/AT-rich regions) may not be well covered by sequence reads. The kmer distribution of *T. socialis* shows a peak of kmer copy number around 40, which does not indicate that a high proportion of kmers have low coverage. However, the effect of sequencing bias cannot be completely discounted and may partly responsible for gap content.

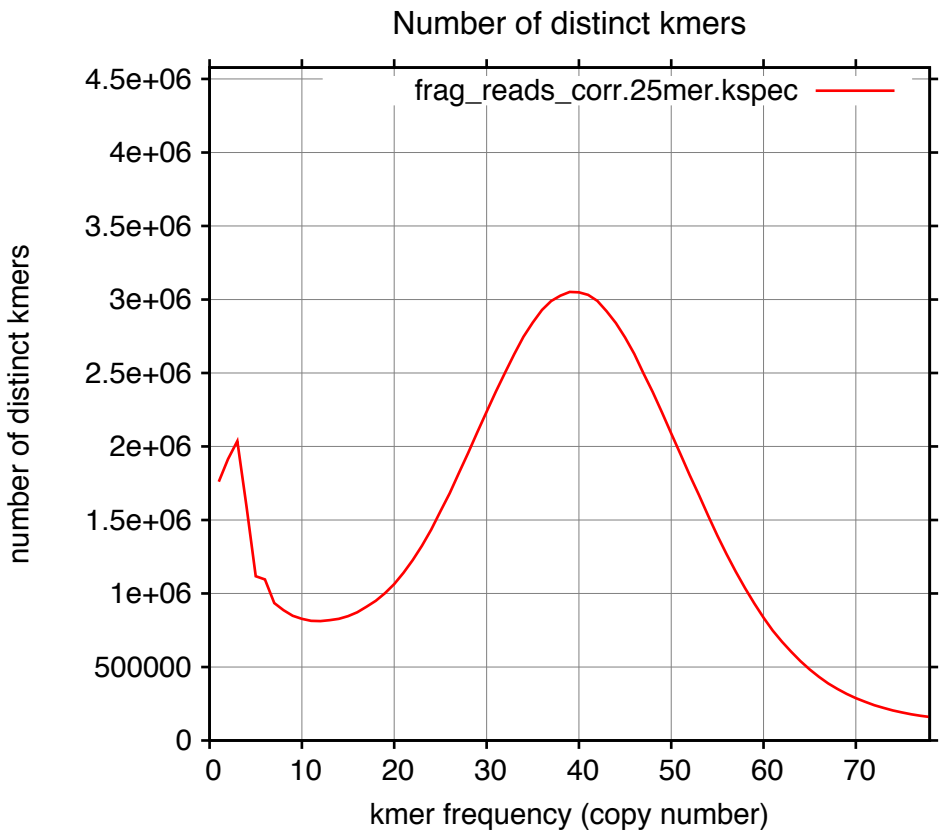


Figure 7.3: Kmer distribution plot generated by Allpaths-LG. A kmer of 25 was used. This figure is useful for the evaluation of GC sequencing bias, ploidy and for the evaluation of data quality. An extensive leftward bias might indicate poor quality input data or GC bias. Typically, a haploid organism produces a single peak as seen in this figure, whereas a diploid organism would present with two similar sized peaks. A peak frequency of 40 copies per kmer is a normal profile for good quality data without extensive GC bias.

A cumulative histogram of kmers, which accumulates from low copy kmers to high copy-number kmers, indicated that approximately 60% of

accumulated kmers possess a low copy number (fewer than 100). The accumulation of low copy-number kmers occurs exponentially while accumulation beyond 60% shows how the high copy-number kmers contribute in a linear fashion towards the cumulative total number of distinct kmers (Figure 7.4). This is suggestive of a high proportion of repetitive sequence content. Based on the 25-mer-kmer distribution Allpaths-LG estimated the genome repeat content to be approximately 43% and estimated the total genome size to be 179,4mb. If correct, this indicates that only ~55% of the genome was assembled in contigs while 77% of the genome was assembled in scaffolds. However, kmer distribution data indicated that missing data are probably highly repetitive and unlikely to include a substantial fraction of coding sequence and result from GenomeScope do not suggest that the genome is this large.

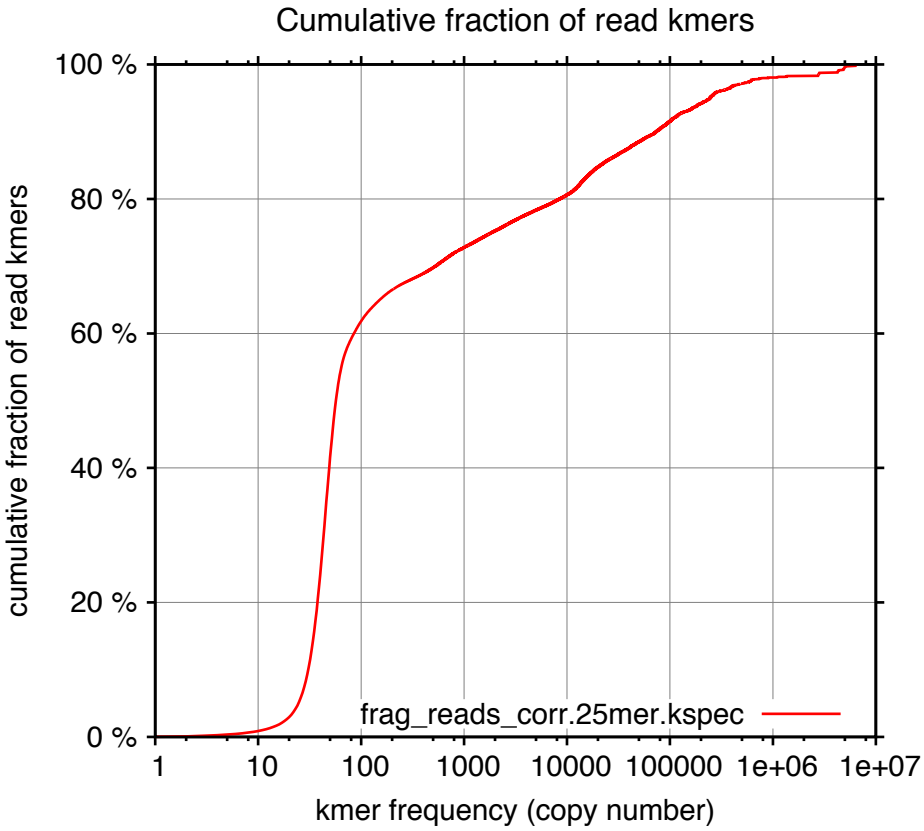


Figure 7.4: Cumulative histogram of 25mer kmers. A cumulative histogram is useful for assessing the degree of genome repetitiveness. Kmers are sorted according to frequency from kmers with a low copy number to kmers of high copy number. The accumulation of kmers shows that approximately 60% of all *Tetrabaena socialis* kmers are accounted for by kmers with fewer than 100 copies. Further accumulation beyond 60% proceeds linearly rather than exponentially. The point at which the accumulation of kmers switches to a linear rate can be used to estimate the genome repetitiveness. In *Tetrabaena socialis* the point is 60%, which indicates an approximate genome repetitiveness of 40%. Allpath-LG calculated the genome to be 43% repetitive.

7.2.5 Genome Coding Content

Both CEGMA and BUSCO were used to assess the coding potential of the *T. socialis* genome as well as other volvocine genomes. Using both methods, the results for *T. socialis* were poor (CEGMA 39% complete and BUSCO 22%). However, the presence and intactness of core eukaryotic genes identified by CEMGA and BUSCO did not correspond with the degree of assembly continuity of the genome assemblies of *C. reinhardtii* (v5), *G. pectorale* or *V. carteri*. In addition, assembly continuity (as reflected by scaffold N50) correlated poorly with the detection of KOGs (r^2 of 0.415 for CEGMA). The high-quality *C. reinhardtii* (version 5) genome was found to possess 90% of KOG genes in complete form, while CEGMA identified 85 and 79% complete KOGs respectively in the genomes of *V. carteri* and *G. pectorale*. By comparison, the genome assemblies of *Steinernema MCB28* and *Rosa spp* were found to possess 93 and 77% complete KOGs- despite scaffold N50 values of 15,8kb and 20.25kb respectively. This disparity is best illustrated by the facts that the *C. reinhardtii* assembly is 417 times more contiguous than that of *Steinernema*, *C. reinhardtii* is essentially resolved to chromosome-level continuity with extensive evidence indicating that most genes are contained in the assembly, and yet fewer complete KOGs were identified in *C. reinhardtii* than in *Steinernema*. Results for BUSCO were even poorer (Table 7.7). Both BUSCO and CEGMA use gene-modelling algorithms to generate KOG gene models. BUSCO utilises augustus

(Stanke, et al. 2006) for gene-modelling and the latest *C. reinhardtii* HMM could be used.

However, even using a well trained HMM for *C. reinhardtii* gene modelling, only 60% of BUSCO eukaryotic genes were identified. These inconsistencies were likely due to the poor representation of KOG genes from GC-rich chlorophytes as most genes used to define core eukaryotic genes are derived from metazoa, yeast and plants, which are typically either balanced or AT-rich. Therefore, further evaluation of coding potential in the *T. socialis* assembly was performed by other means. RNAseq reads were mapped to the *T. socialis* genome and 79% of paired-reads were identified to map to the genome as pairs. Furthermore, 92% (21 823) of transcripts from the *T. socialis de novo* transcriptome (23 559 transcripts) aligned to the *T. socialis* genome (all with greater than 82% percentage identity and 90% > 95% percentage identity). A total of 93.8% of KOG domains (n=458) were identified within the MAKER2 derived *T. socialis* proteome (HMMER $E=1x10^{-5}$). By comparison, scanning of KOG domains against the proteomes of *V. carteri* (v1) and *C. reinhardtii* (v5.5) identified 96,3% and 98,7% domains respectively. Complete and manually curated *C. reinhardtii* KOG orthologues (n=406) were obtained from the CEGMA website (http://korflab.ucdavis.edu/datasets/cegma/). *Tetrabaena socialis* proteins (from MAKER2 annotations) were aligned to *C. reinhardtii* KOG proteins using blastp and subsequently sorted to retain only the highest scoring matching *T. socialis* homologues.

A total of 382 proteins (94%) were identified with homology ($E=1x10^{-10}$) to *C. reinhardtii* KOG proteins, of which, a total of 367 unique *T. socialis* proteins (90%) were identified. 92% of filtered *T. socialis* homologues of *C. reinhardtii* KOG proteins aligned to their homologous partners with greater than 50% while 79% aligned with greater than 70% (n=290). A total of 393 (96.7%) *C. reinhardtii* KOG proteins aligned to the *T. socialis* genome ($E=1x10^{-10}$). The alignment of each KOG gene to the *T. socialis* genome was not manually inspected and the 96.7% of KOG genes that mapped to the *T. socialis* genome includes only the top high-scoring segment pair (HSP) for each query KOG protein (i.e. the blast option max_target_seq=1 was used). Together, KOG domain content, RNAseq mapping, transcriptome mapping and *C. reinhardtii* KOG protein alignment to the genome and predicted proteome of *T. socialis* indicate that between 79 and 96% of genes were present in the *T. socialis* assembly and that possibly up to 90% of genes were modelled by MAKER2. However, in this comparison the difference between annotation completeness versus genome-assembly completeness must be considered as genes might be present within the assembly but un-annotated, genes may have been truncated/erroneously modelled and biological differences between *T. socialis* and *C. reinhardtii* proteins may also affect alignment rates. Undoubtedly, the genome contains more coding-content than the predicted proteome, however, this difference is not exactly known but based on the presence and completeness of, for

example, KOG genes missing genes are expected to be comparable to other draft genome assemblies.

Table 7.7: CEGMA results for volvocine genomes as well as the genomes of *Steinernema MCB28* and *Rosa spp*.

Organism	Scaffold N50	CEGMA Complete%	BUSCO Complete%
***Chlamydomonas reinhardtii* v5**	6.617mb	90.73	60
***Volvox carteri* v2**	2.599mb	85.08	43
***Gonium pectorale* v1**	1.267mb	79.44	54
Pleodorina starrii	12.98kb	52.82	23
Rosa spp	20.25kb	79.84	51
Steinernema MCB28	15.84KB	93.95	57
Tetrabaena socialis	145.9kb	39.11	22

7.2.6 Genome Contamination

Complete cpDNA and mtDNA molecules were not identified in the final nuclear genome assembly, whereas a single contig containing the Illumina “spike-in” control genome, lambda *phiX174*, was identified and removed. Smaller elements of organelle genomes could be found within the nuclear genome assembly. It is unknown if these elements represent intracellular horizontal transfer events or were erroneously incorporated into the assembly. The level of organelle presence in the nuclear genome of *T. socialis* was compared to the level in *C. reinhardtii* and found to be similar. The longest matching mitochondrial element in the nuclear

genome was 714bp and mtDNA segments were average of 76,3bp in length. The longest alignment of cpDNA to the nuclear genome was 4096bp and an average of 75,3bp.

The total fraction of nuclear regions with homology to organelle DNA was less than 0,3%. Generating entirely axenic cultures of algae can be challenging and hence DNA from *T. socialis* was expected to contain a degree of contaminating bacterial or viral elements. A previous analysis that assessed contaminants in several publically available eukaryotic genomes using Kraken identified 227 *C. reinhardtii* contigs to possess contaminants. Kraken searches against the genomes of *G. pectorale* and *T. socialis* identified that as many as 14% of scaffolds contain evidence of bacterial of viral contamination. These results, however, are purely based on the presence of short kmers and thus constitute only a small fraction of the genome (for *C. reinhardtii* <0,25%). Overall, the total contaminating sequence (organelle and foreign) in the nuclear genome assembly of *T. socialis* is estimated to be between 0.3-0.6%.

7.2.7 *De novo* transcriptome

7.2.7.1 Trimming

RNAseq data was of high sequence quality, although up to ~35% of reads contained adaptor sequences (towards the end of reads). In the author's experience this is typical of RNAseq. Using the TransRate assembly score as a metric for evaluating transcriptome quality a study

that examined 155 published *de novo* transcriptomes identified that read quality was the major determinant of *de novo* transcriptome quality.

7.2.7.2 Transcriptome assembly results

A kmer of 25 produced the best assembly for Bridger and Trinity 2 performed better than Trinity version 1 (Trinity only performs assembly with a fixed kmer of 25). Although Trinity is probably the most commonly used assembler for *de novo* transcriptomes TransRate indicated that Bridger *de novo* produced a superior assembly. This was reflected in mean contig length, number of contigs over 1kb in length, the percentage of 'good mappings' (paired-end reads that map concordantly) as well as cumulative assembly metrics such as N30, N50 and N90. In favour of Trinity 2 over Bridger *de novo* assemblies were the higher fraction of identifiable open reading frames (ORFs - 76,5%) and consequent higher percentage of *C. reinhardtii* reciprocal blast orthologues (Table 7.8). However, for downstream annotation applications the increased length of transcripts and higher accuracy (as reflected by 'good mappings') produced by Bridger were considered to outweigh the superior ORF percentage of Trinity 2.

Table 7.8: Summary of transcriptome assembly results. The table includes summary statistics and quality metrics generated with TransRate.

	Bridger *de novo* (k=25)	Bridger *de novo* (k=31)	Trinity v1 (k=25)	Trinity v2 (k=25)
%GC	64	64	64	64
No. of contigs	27 712	27 384	29 255	34 908
Optimal Assembly Score	0.362	0.358	0.354	0.357
% rbh per reference	46.4	45.2	46.99	52.8
Mean ORF %	72.9	73.6	75.8	76.5
Good Mappings %	78.8	78.7	75.0	74.9
N30	1 055bp	1 016bp	888bp	856bp
N50	670bp	637bp	572bp	554bp
N90	414bp	244bp	241bp	241bp
Mean Length	525bp	510bp	478bp	471bp
No. over 1kb	3 259	2 957	2 519	2 791
Unique "Genes"	24 791	25 204	24 508	28 801
No. of discarded transcripts	4 153	3 897	5 161	11 251

Approximately 50% of 155 published transcriptomes have a TransRate score lower than 0.35 and therefore the Bridger *de novo T. socialis* assembly may be considered superior to over 50% of these assemblies.

7.3 Discussion

7.3.1 Nuclear Genome Quality

There remains no definitive means to assess the quality of a genome assembly and its associated annotation. Computational methods used to generate both the assembly and annotations are likely to be replete with errors. This is particularly true for genomes assembled using “short read” NGS technology. Indicators of assembly continuity such as the number of scaffolds or contigs, scaffold or contig N50, the length of largest scaffold and the percentage of gaps are typically reported for genome assemblies. They do not necessarily correlate with the coding content of the assembly. Protein coding regions of the genome are typically the genomic content of interest and thus objective means to evaluate the coding-potential of a *de novo* genome assembly can be useful for assessing genome quality in terms of coding potential. Programs such as CEGMA and BUSCO are common tools that in many lineages have proven of useful for assessing the coding potential of a *de novo* assembly. However, the robustness of these algorithms across diverse eukaryota has not necessarily been fully assessed. As with many aspects of biology such algorithms and the conserved content they include tends to be well suited for animal and plant lineages. CEGMA and BUSCO results for volvocine algae were demonstrated to be incongruent with assembly continuity.

However, it must be noted that the more fragmented the volvocine genome assembly, the poorer the results from CEGMA and BUSCO were, even though the correlation between assembly continuity and CEGMA results was poor (r^2 of 0.415). The presence of BUSCO conserved eukaryotic genes in *C. reinhardtii* was remarkably low when the quality of the genome assembly is considered. Version 5 of the *C. reinhardtii* assembly has almost been entirely resolved into chromosomes and it is highly unlikely that 40% of genes are absent from the assembly. The extent of investment into *C. reinhardtii* transcriptome analysis as well as EST sequencing supports the above statement (Blaby, et al. 2014). Even the ~10% of missing CEGMA genes is unlikely to be a true reflection of missing coding sequence in the *C. reinhardtii* assembly. It is not entirely known why CEGMA and BUSCO did not perform well on volvocine genomes. Some possible reasons might include the high GC content, the poor representation of algal genome sequences in databases used to build markov models of core eukaryotic genes and it is also possible that certain genes deemed to be universal in eukaryotes are not present in the volvocines. Unfortunately, the inadequate performance of BUSCO and CEGMA indicated that they could not be considered as reliable indicators of genome assembly quality for the de novo genome assembly of *T. socialis*. Alternative measures were taken to address the coding content of the *T. socialis* assembly. These measures included assessments of both the genome and predicted proteome of *T. socialis*.

Although the CEGMA and BUSCO algorithms did not perform well on volvocine genomes at building KOG genes *ab initio* with HMM models, a previously annotated list of *C. reinhardtii* KOG genes (n=406) was used to assess the *T. socialis* genome and proteome. More putative KOG orthologues were identified through aligning proteins from this list to the *T. socialis* genome and predicted proteins than what either CEGMA or BUSCO identified. Alignments of *C. reinhardtii* KOG genes to the *T. socialis* genome and predicted proteome respectively identified 96,7% and 90,4% of genes to be present. These results indicate that the number of genes missing not modelled by MAKER2 due to insufficient evidence or other factors but are present in the genome is not likely to be significantly greater than ~6%. In the CEGMA implementation a standard has been applied that if 70% or greater of a gene model was identified then the gene was considered complete. It is unclear if this standard has any relevance to blast based alignments. However, if these standards are used then 71 % of proteins in the *T. socialis* predicted proteome might be considered to complete. DOGMA results indicate a similar pattern (67% complete), however, *C. reinhardtii* v5.5. annotations are only 90% complete according to DOGMA. *Arabidopsis thaliana* was the only photosynthetic organism included during development of DOGMA domains.

According to Allpaths-LG kmer analysis the genome of *T. socialis* is nearly 180mb in length and over 50% repetitive. If correct, the increased size of the *T. socialis* nuclear genome is unique amongst sequenced

volvocines and is a greater expansion of nuclear genome than seen in other more complex volvocine algae. Based on volvocine phylogeny the expansion of the *T. socialis* nuclear genome is likely to have occurred after the divergence from the Goniaceae and the Volvocaceae and is therefore likely, but not definitively, to be more recent in origin. Genomes that are replete with highly similar repetitive sequences that arose relative recently are notoriously difficult to assembly (Kovach, et al. 2010). Active and rapid expansion of repetitive elements throughout a genome results in a high level of sequence similarity amongst repetitive elements, which greatly hinders assemblers that are readily stymied by close sequence similarity. The similarity of repeat sequences was not assessed and most of these repeats will not have been assembled. Factors that influence genome assembly continuity include the percentage of repeat content, the similarity of repeat content, GC content and total genome size (Nagarajan and Pop 2013). A genome assembly with a substantial gap content (denoted by N's in the sequence data) and missing data does not necessarily indicate that the coding content of the genome is poor or at that useful conclusions cannot be drawn from such an assembly. For example, the genome of *Coffea canephora* was assembled to 80% of the estimated genome size and possessed 17% gap content (essentially only 66% of the genome was assembled) but was non-the-less useful for exploring the origin of caffeine biosynthesis – a topic that most scientists would find relevant (Denoeud, et al. 2014).

However, the greatest challenge with missing data is to “prove a negative” i.e. if a particular gene is missing, is this due to the region being missing from the assembly (a false negative) or a real loss (true negative). In the age of draft genome assemblies missing genomic regions are likely to be commonplace and the key is to maintain rigorous in the interpretation of draft assemblies. Evaluation of sequence reads indicated that in terms of GC content that the genome of *T. socialis* was similar to *G. pectorale* and *C. reinhardtii* (~65% GC). High GC and repeat content hindered the generation of a high quality draft genome assembly and thus a series of assembly algorithms were evaluated. Genome assembly occurred through an iterative process of generating sequence data, testing various assembly algorithms and then implementing new sequencing and assembly strategies. Initially, a PE library of 300-600bp insert was sequenced and assemblers were performed (Allpaths-LG could not be tested with these data as they did not significantly overlap and overlapping reads are a requirement for Allpaths-LG). Gel-free mate-pair data were obtained and scaffolding was attempted but gel-free data were found to be inffective for scaffolding and therefore a size-selected MP library was generated and sequenced.

Although the inclusion of size-selected data did improve scaffolding high-quality assemblies could not be produced. To establish the optimal strategy for further sequencing simulated data were generated from the *C. reinhardtii* reference genome (v5) using ART_Illumina (Huang, et al.

2012). Simulated data of 100X coverage was generated for three separate libraries (Overlap PE data, Miseq data and MP data).

The assemblers CLC, SGA, JR Assembler, Allpaths-LG and DISCOVAR de novo were tested. Scaffolding of DISCOVAR contigs was performed using SSPACE, as DISCOVAR did not support scaffolding. Although the contig level statistics for DISCOVAR de novo were superior to Allpaths-LG independent scaffolding algorithms could not obtain the scaffold level contiguity that Allpaths-LG produced. Therefore, an overlap library for *T. socialis* with an insert size of ~180bp was created and sequenced. Later, Miseq data with longer reads were generated to further improve the assembly. For *T. socialis*, the high repetitive sequence content, the high GC content and the possible recent origin of repetitive sequence hindered assembly with short-read NGS data. Further efforts to improve the sequence assembly will require long read technologies such as those obtainable with Oxford Nanopore, Pacific BioScience and Illumina TruSeq Synthetic Long-Read sequencing.

7.3.2 Transcriptome

7.3.2.1 General

The aim of generating a *de novo* transcriptome for *T. socialis* was to supply organism specific transcript/gene evidence for genome annotation pipelines. Furthermore, a *de novo* transcriptome was generated to explore genes missing from the *de novo* genome and thus

genome-guided transcriptome assemblies were not used. Numerous options for *de novo* transcriptome generation are available to further optimise transcriptome results.

For the purpose of version 1 of the *T. socialis* genome and annotation a single "best transcriptome assembly" was generated and utilised for gene annotation. The *de novo* transcriptome assemblers Trinity (version1 and 2) and Bridger *de novo* were evaluated. Both assemblers generated the best assemblies with a kmer of 25 and Bridger *de novo* produced superior assemblies than Trinity at both kmers of 25 and 31. A lower kmer was not utilised because lower kmers may be error-prone. The 25mer assembly of Bridger *de novo* was selected for annotation purposes.

7.3.2.2 Quality Assessment of Bridger De Novo 25mer Assembly

The TransRate optimal assembly score of the selected 25mer Bridger de novo assembly of 0,362 was compared with a published dataset of TransRate optimal assembly scores that were generated for 155 published de novo transcriptomes. This score corresponds with an average optimal assembly score- an optimal assembly score of 0,35 is better than 50% of the 155 published de novo transcriptomes (Smith-Unna, et al. 2016). The number of transcripts (n=27 712) is somewhat low for a *de novo* transcriptome assembly, which are prone to generating many hundreds of thousands of transcripts. This may in part be due to the

use of a single time point in the life cycle of *T. socialis*. It is also probable that, as with the genome assembly, the high GC content of coding regions may have affected the transcriptome assembly. Further improvement of the *T. socialis* transcriptome would be to generate multiple RNAseq libraries that derive from multiple points in the full life cycle. mRNA abundance fluctuates throughout the life and certain genes may only ever be expressed during certain phases of the life cycle or in specific cell types/tissues. Expression data from multiple time points would improve the number of genes captured in the de novo transcriptome assembly.

8. APPENDIX B: NUCLEAR GENOME ANNOTATION

8.1 Methods

8.1.1 Annotation of Gene models

Details of MAKER2 gene modelling are described in Chapter 3 (Featherston et, 2018). Briefly, MAKER2 (Holt and Yandell 2011) gene models were derived from numerous sources of evidence and the best gene model was computed from 3 *ab initio* gene-modelling algorithms that were informed by evidence. An example of a MAKER2 gene model viewed using the Apollo genome browser (Lewis, et al. 2002) can be seen in Figure 8.1. A total of 33 030 unfiltered (for AED or domains) gene models, were generated by MAKER2, of which, 10 734 (32.5%) had AED scores <1. After filtering to retain only gene models with an AED less <1 or with a domain 15 513 gene models survived. Therefore, a total of 4779 (30.8%) gene models were retained that were derived exclusively from *ab initio* predictions. A total of 303 226 protein homology matches and 44 197 transcript matches were used as hints and evidence for gene finders within the MAKER2 environment. A series of annotation quality metrics were measured. Annotations were examined for domain content (including Eukaryotic Clusters of Orthologous Groups (KOG) domains), the number of volvocine orthologues (using OrthoMCL- Chapter 3 - and visualised using OrthoVENN), the number of proteins with significant homology to the nucleotide resource (NR) database, the number of proteins with homology to other volvocine proteomes and the distribution of AED scores.

Quality metrics based on functional annotations are described in further detail under the heading Functional Annotation. Summary statistics were generated with a perl script “GxF_to_stats.pl” that was obtained from Github (https://github.com/EVidenceModeler/EVidenceModeler/blob/master/EvmUtils). A cumulative histogram plot of AED scores was generated using the MAKER2 accessory script “AED_cdf_generator.pl” and was drawn in R.

8.1.2 Functional Annotations

Blast based techniques, HMMER domain scans and InterProScan scans are described in Chapter 3. The Mercator pipeline for plant proteome annotation was accessed via the web portal http://www.plabipd.de/portal/mercator-sequence-annotation and was used to assign MapMan annotations to the proteomes of *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* (Thimm, et al. 2004). MapMan annotations included functional descriptions and identifiers that like, GO terms, were stratified by degree of specificity.

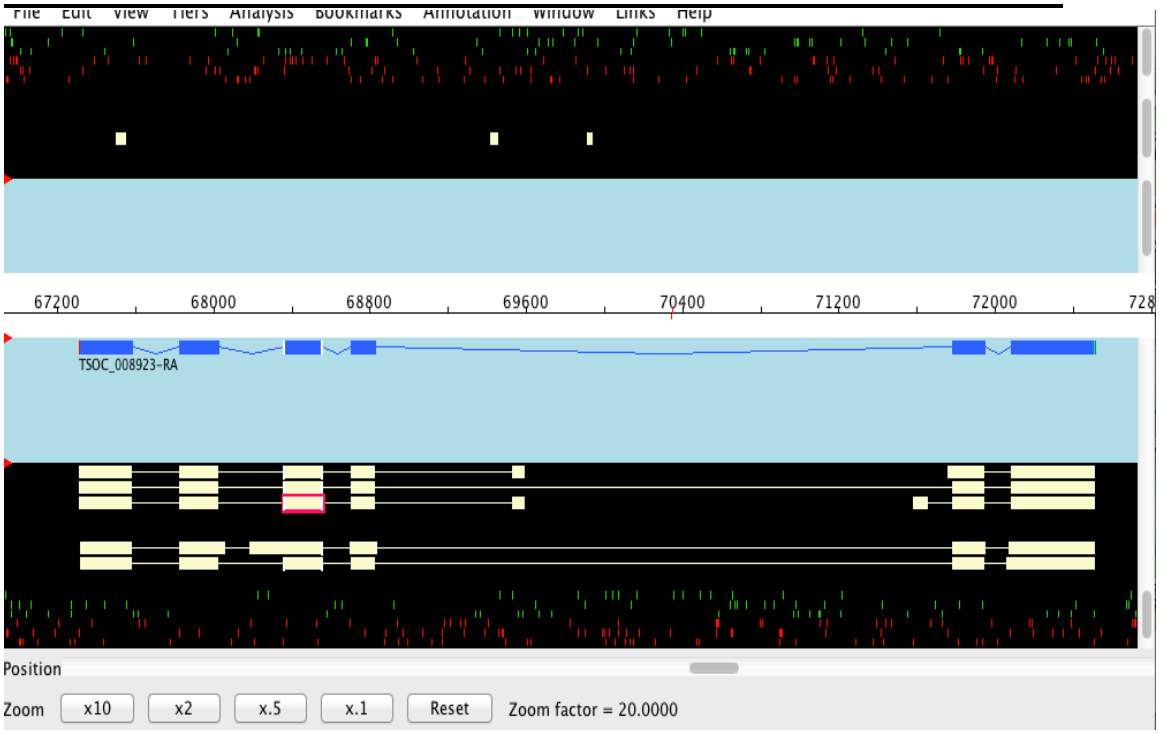


Figure 8.1: A screenshot of *Tetrabaena socialis* MAKER2 gene model TSOC_008923-RA as viewed with Apollo genome browser. In blue is the consensus MAKER2 gene model. Shaded blue regions indicate exons while connecting lines represent introns. The first three lines seen below the MAKER2 gene model are gene models produced by augustus (Stanke, et al. 2006), Genemark (Borodovsky and Lomsadze 2011) and SNAP (Korf 2004). Lines 4 and 5 below the MAKER2 model are from protein alignments of *Chlamydomonas reinhardtii*, *Gonium pectorale* of *Volvox carteri* proteins. For this model the Genemark model was closest to the protein homology based evidence. The annotation-edit distance (AED) of this model was 0.08 and can thus be considered as a model that closely matches available evidence. Above the MAKER2 model are 3 regions masked by repeat masker can be seen.

8.2 Results

8.2.1 Gene models

A cumulative plot of AED scores against the number of proteins indicated that ~86% of proteins with AED scores <1 had an AED score of 0.5 of lower (Figure 8.2). A summary of annotation metrics in volvocine algae is provided in Table 8.1.

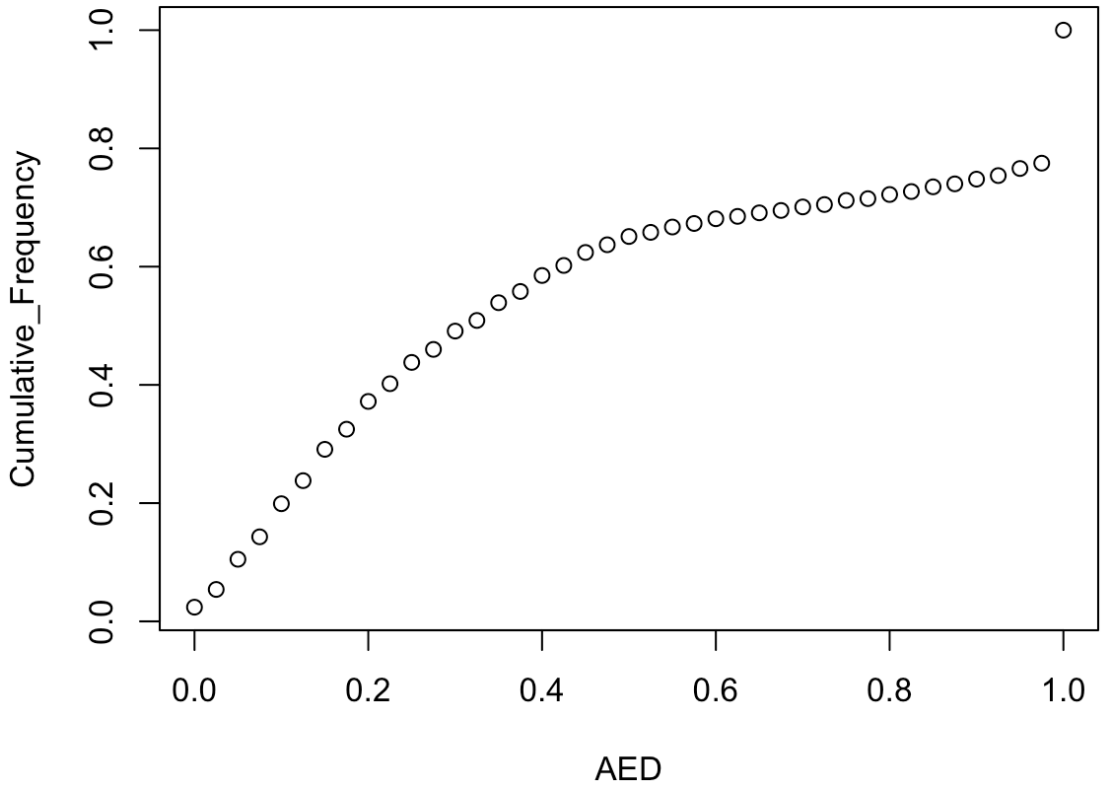


Figure 8.2: Cumulative histogram of annotation edit distance scores.

X axis: annotation edit distance score for MAKER2 gene models.

From left to right the AED increases and therefore the quality of gene models decreases due to lower levels of evidence or greater levels of disparity between evidence and the computationally derived gene model. An AED of 1 indicates that a gene model was generated without any evidence by *ab initio* gene prediction algorithms. Gene models with an AED of 1 were retained if a domain could be found in the protein product of the gene model. MAKER2 derived genes were evaluated for repeat content. The intersection of regions masked by RepeatMasker (using a library of *V. carteri* repeats) (Tempel 2012) with MAKER2 genes (total gene span including introns) was evaluated at the level of 30 and 70% overlap (or intersection). A total of 94 genes were identified that overlap by 30% or greater with repeat masked regions whereas 22 genes were identified that overlap by 70% or greater. The approach was repeated for *G. pectorale* and 847 genes were found to overlap with repeat masked genome regions by 70% or greater. The total number genes that house a substantial portion of repetitive content was considered minimal. MAKER2 is likely to have "over-masked" genes as opposed to having generated a substantial repertoire of spurious genes derived from repetitive elements.

Table 8.1: Summary Statistics of Volvocine Annotations

	Tetrabaena socialis	*Volvox carteri* v1	*Chlamydomonas reinhardtii* 5.5	*Gonium pectorale*
Genes	15 513	14 520	17 737	17 984
Exons per gene	5.8	7.78	7.46	6.94
Avg Gene length	3 752	5 269	5 342	3 925
Avg Exon length	223	194	255	210
Avg Intron Length	635	496	279	349
Avg Protein Length	411	568	464	487
Complete proteins (stop codons)	12 983	13 134	19 501 (including alternative splices)	Undefined
Domain Presence (no of proteins)	9 803 (63%)	8 456	10 889 (including alternative splices)	10 029
Total domains	14 041	13 160	16 200 (including alternative splices)	15 786
AED of 1	4 779 (31 %)	NA	NA	NA
Unique Domains	2 998	3 269	3 482	3 340

Note* Gene length includes UTR regions.

Reciprocal blast hits were used to identify orthologues between volvocine proteomes (see Chapter 3 for methods using OrthoMCL). Fewer orthologues were identified for *T. socialis* proteins than for other volvocines (Table 8.2). However, reciprocal blast orthologues between multicellular proteomes and the high quality *C. reinhardtii* proteome

showed that the percentage of *T. socialis* proteins with orthologues to *C. reinhardtii* is similar to those of *G. pectorale* and *V. carteri*.

Table 8.2: Orthologues between Volvocine Proteomes.

	T. socialis	***C. reinahardtii***	***G. pectorale***	***V. carteri***
T. socialis	X	8 931	6 730	6 569
C. reinhardtii	8 931	X	10 047	9 384
G. pectorale	6 730	10 047	X	8 417
V. carteri	6 569	9 384	8 417	X

DOGMA (Dohmen, et al. 2016) was used to assess the completeness of the volvocine proteomes by examining core domains (derived from *A. thaliana*, *C. elegans*, *Mus musculus*, *H. sapiens*, *D. melanogaster* and *S. cerevisiae*) and the arrangement of multi-domain proteins. As per CEGMA and BUSCO genome assessments, volvocine proteomes were found to be missing many domains (Table 8.3). The domain content of *T. socialis* was found to be 58.5%. The first core set of 965 conserved domain arrangements (CDA) performed best for *T. socialis* and *G. pectorale*. Of 317 CDAs absent from the *T. socialis* proteome 87 were also absent from *C. reinhardtii*. If these CDAs are excluded the *T. socialis* proteome is 76% complete.

Table 8.3: DOGMA Assessment of Proteome completeness for Volvocine Algae.

Species	Core set 1	Core set 2	Core set3	Total
Chlamydomonas reinhardtii	89.37	87.72	87.12	88.75
Tetrabaena socialis	67.91	41.18	39.39	58.50
Gonium pectorale	83.40	75.70	62.88	79.62
Volvox carteri	82.59	78.26	83.33	81.54

8.2.2 Functional annotation

The outcome of homology searches against other volvocine proteomes as well as against the UNIPROT databases trEMBL and SWISSPROT (Boutet, et al. 2007) are summarised in Table 8.4. InterProScan identified 8072 GO terms, 9803 proteins were found to contain PFAM-A domains and 14 041 total domains were identified of which 2998 were unique. These results are comparable to *C. reinhardtii* annotations as well as those of *V. carteri* and *G. pectorale*. Overall, MapMan annotations were assigned to very few volvocine proteins (~30% of proteins). The terms were, however, explored for changes between *C. reinhardtii* and the multicellular taxa. A summary of identified MapMan terms for *T. socialis* proteins is found in Figure 8.3. The number of un-assigned proteins is somewhat larger in *T. socialis* than in other volvocines (*T. socialis* 73.2; *V. carteri* 69.72; *G. pectorale* 69.74; *C. reinhardtii* 69.68).

Table 8.4: Blast results of *Tetrabaena Socialis* Proteins against UNIPROT and the Proteomes of other Volvocines. Note that results include only top blast hits ($E=1x10^{-7}$).

Query Proteomes	*Tetrabaena socialis* matches (15 153 proteins)
C. reinhardtii	12 954 (83.5%)
G. pectoral	12 740 (82.1%)
V. carteri	12 218 (78.7%)
Volvocines	13 712 (88.3)
UNIPROT-SWISSPROT	7 838 (50.5)
UNIPROT-TREMBL	13 091 (84.3%)
All Blast hits	13 982 (90%)

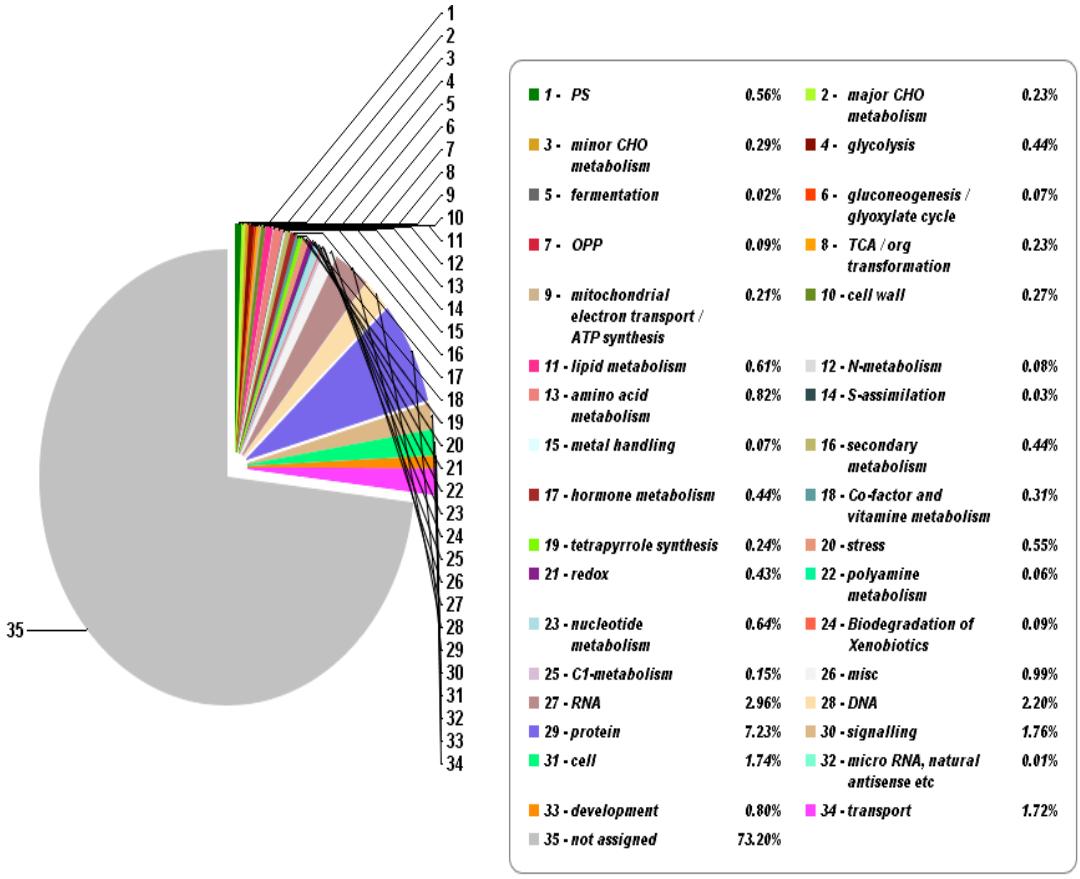


Figure 8.3: Pie chart of MapMan terms assigned to *Tetrabaena socialis* proteins.

8.3 Discussion

8.3.1 Gene Models and Summary Statistics

The total of 15 513 gene models is consistent with other volvocine algae (Table 8.1). The number of genes without evidence (31%) is lower than that of *V. carteri* v1, where 67% of genes were generated exclusively using *ab initio* methods, however evidence for later versions of *C. reinhardtii* (e.g. version 5.5) and for *G. pectorale* exceeds that of *T. socialis*.

The total gene number of *T. socialis* and *V. carteri* are closer to one another than to *C. reinhardtii* and *G. pectorale*. *Volvox carteri* v1 annotations like those of *T. socialis* were generated using a pipeline that included both protein homology and assembled EST evidence whereas an alternative strategy was followed for the later *C. reinhardtii* and *G. pectorale* annotations. A number of possible strategies could have been followed to generate annotations for *T. socialis*. RNAseq data mapped to genomic sequences was the predominant technique used to generate the newer versions of *C. reinhardtii* annotations and the *G. pectorale* v1 annotations. While this technique when used with a large dataset of life cycle RNAseq data may detect more genes than the method used to generate the *T. socialis* annotations, it is also likely to produce a higher level of erroneous fusion of adjacent coding regions into longer models than to first assemble RNA reads into contigs and use the contigs as EST evidence (The erroneous fusion of adjacent coding regions by using

mapped RNAseq reads to generate hints for gene finders can be mitigated by iteratively mapping reads to potential exon-exon junctions predicted by gene finders). Although RNAseq derived data is invaluable for annotation the inclusion of protein homology data and the combination of multiple gene finder predictions can assist with mitigating the errors that can result from the exclusive use of any single line of evidence (Yandell and Ence 2012). Thus *T. socialis* annotations were performed using *de novo* assembled transcripts, protein homology data and three gene finder algorithms within the MAKER2 pipeline.

Although accuracy may have been improved using the MAKER2 pipeline the number of genes for *T. socialis* may be lower than if RNAseq data from a 24-life cycle was used to generate assembled ESTs or if the same procedure used to annotate the *C. reinhardtii* and *G. pectorale* genomes was used. It is unclear as yet if the number of genes in *V. carteri* is truly lower than in *C. reinhardtii* and *G. pectorale* and if the number of genes in *T. socialis* should be higher. *Chlamydomonas reinhardtii* annotations have increased from ~15 377 to ~17 713 protein coding loci between v3 and v5.5 while *V. carteri* annotation from v1 to v2.1 have remained stable at approximately ~14 500 (14 520 and 14 237 respectively) protein coding loci. Given the current state of annotation methodology the number of protein coding loci may vary significantly between version updates. It is noted that as per any draft genome assembly missing genes from *T. socialis* could be due to either insufficient

evidence for annotation or gaps in the genome assembly. The average gene length and protein length is statistically lower in *T. socialis* than in other Volvocines. Further improvements in annotation and genome quality are likely to alter these metrics. However, the present annotation contains a similar number of domains.

9. APPENDIX C: ADDITIONAL COMPARATIVE GENOMIC ANALYSES

9.1 Methods

9.1.1 Syntenic re-arrangements

SyMap (Soderlund, et al. 2006) was used to evaluate the degree of synteny between *T. socialis* and *C. reinhardtii*. Given the fragmented nature of the *T. socialis* assembly only scaffolds greater than 100kb in length were included. Both dot plots and circular images of synteny were generated. The fragmentation of the *T. socialis* assembly limited an accurate assessment of synteny and hence the *T. socialis* assembly was not compared to the *G. pectorale* and *V. carteri* assemblies.

9.1.2 Phylogenomics

As per described in Chapter 3 a phylogenomic "supermatrix" protein phylogeny (Delsuc, et al. 2005) approach was followed to evaluate the taxonomy of whole-genome sequenced chlorophytes. The phylogeny in Chapter 3 was extended to include selected transcriptome sequenced chlorophytes. Proteome data were obtained from sources summarised in Featherston et al (2018). Proteomes for whole-genome sequenced chlorophytes as well as transcriptome-derived proteomes belonging to taxa from the *C. reinhardtii* order were included. Transcriptome-derived proteomes were obtained from the Marine Micro-Algae Transcriptome Repository, which houses thousands of transcriptomes.

Proteomes derived from transcriptome assemblies were limited to species belonging to the Chlamydomonadales order because these species are most related to the volvocine lineage. A total of 18 proteomes were included for the phylogeny. Splice variants from the *C. reinhardtii* v5.5 proteome were removed to include only the primary protein. Putative orthologues and orthologue families were identified using OrthoMCL version 2.09 (Li, et al. 2003; Fischer, et al. 2011). A markov clustering (MCL) inflation value of 1.5 and an E-value of $1x10^{-5}$ were used as OrthoMCL parameters. After MCL clustering a perl script was used to remove families that contained more than 1 protein per species (paralogues) as well as to remove families that did not include proteins from all 18 species (https://github.com/hyphaltip/genome-scripts/). Thus, the filtered OrthoMCL families each possessed 18 proteins and each protein was from a different species (these are known as singletons). Proteins from each family were separately aligned using MUSCLE (3.8.31) (Edgar 2004) and alignments were trimmed with trimal (1.3) (Capella-Gutierrez, et al. 2009). Trimmed alignments were concatenated using fastconcat (Kuck and Meusemann 2010). ProtTest (v3) was used to select the most appropriate phylogenetic model of amino acid substitution (Abascal, et al. 2005).

The Dayhoff, WAG, JTT, VT, Blosum62 and DCMut matrices as well as all available distributions were tested and evaluated based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

and the Decision Theory Criterion Framework (DT). RAxML (8.2.0) (Stamatakis 2014) was used to generate a phylogeny with 100 bootstraps using the model parameters of “gamma”, “invariable sites”, “WAG” and “empirical frequency estimates”.

9.1.3 Comparative genomics: general functional annotations

9.1.3.1 Differences in general functional annotations

To gain insight into functional differences associated with the transition from unicellularity to coloniality/multicellularity in the volvocine lineage, certain functional annotations assigned to the proteins from the multicellular volvocines were compared with those assigned to *C. reinhardtii* proteins. Annotations that were compared were MapMan terms, InterProScan gene ontology (GO) terms and domains identified by HMMER scanning against the Pfam-A database (domain analysis is described in Chapter 3). The above-mentioned comparisons were used to describe “broad” differences in annotations and to include as many “broad” differences as possible false-discovery rate corrections were not applied to statistical tests.

9.1.3.2 Mercator and Gene Ontology term analysis

A χ^2 distribution test was used identify assigned MapMan annotations that were statistically differently (p=0.05) represented between the proteomes of *C. reinhardtii* and the combined proteomes of multicelled

taxa *T. socialis*, *G. pectorale* and *V. carteri.* The same approach was followed for GO terms. GO terms were extracted from InterProScan annotations assigned to volvocine proteomes (i.e. *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri*). A χ^2 distribution test was used to compare GO terms assigned to the proteome of *C. reinhardtii* with a combined list of proteins from the multicelled taxa (*T. socialis*, *G. pectorale* and *V. carteri*).

9.1.3.3 Volvocine genes of interest

9.1.3.3.1 Prochnik table genes

The supplementary information content of the *V. carteri* genome publication included a table of "Proteins involved in processes that are associated with increased developmental complexity in *Volvox* relative to *Chlamydomonas*" (Prochnik, et al. 2010). The table is hereafter known as the "Prochnik table". The table has been updated to include newer versions of *C. reinhardtii* and *V. carteri* proteomes as well as the proteome of *G. pectorale* (Hanschen et al., 2016). Results from various informatic tools, were inspected to identify putative *T. socialis* orthologues of Prochnik table proteins.

Orthologues from *C. reinhardtii, G. pectorale* and *V. carteri* were aligned to the *T. socialis* genome and proteome using tblastn and blastp respectively. Furthermore, OrthoMCL protein-family clusters were inspected and in a few instances protein phylogenies were generated to

identify *T. socialis* orthologues. Further efforts were taken to identify and classify, InvA/IAR1, matrix metalloproteases, pherophorins and cell cycle proteins and are discussed separately below.

9.1.3.3.2 Inversion A (*inva/iar1*)

In *C. reinhardtii*, *G. pectorale* and *V. carteri* the *inva*/*IAR1* gene spans overs 7kb and contains many exons (e.g. 23 in *C. reinhardtii*). The InvA/IAR1 protein was not identified in the *T. socialis* proteome. InvA/IAR1 proteins from *G. pectorale*, *V. carteri* and *C. reinhardtii* were aligned to the *T. socialis* genome using tblastn. However, only a fragment corresponding to the N-terminal of *C. reinhardtii* IAR1 could be identified on scaffold 562. The remaining coding sequence could not be identified in the *T. socialis* genome assembly. Next, the *T. socialis de novo* transcriptome (generated using Bridger) was searched and a single contig was identified that corresponded to *C. reinhardtii* exons 19-21. To improve the completeness of the *T. socialis* protein RNAseq reads were aligned to the Bridger *de novo* derived transcript and extension of the contig was performed by manually extending reads that aligned to the 3' and 5' end of the transcript. This was performed using the "Map Reads To Reference" function within the CLC Genomics Workbench (v8). This, however, was unsuccessful. In addition, the program Price (Ruby, et al. 2013) was also unsuccessful at extending the contig.

RNAseq and PE genomics reads were subsequently mapped to *C. reinhardtii* and *G. pectorale* CDS and gene sequences using the CLC read mapping function with mapping parameters of 90% mapping (90% of the read had to align) and 90% similarity. The consensus sequences generated from reads that mapped to CDS and/or gene sequences were then extracted, translated into amino acid sequences (3 frames) and aligned to the *G. pectorale*, *V. carteri* and *C. reinhardtii* INVA/IAR1 protein sequences to identify the correct translation-frame. However, each region of alignment was manually inspected and for most regions alternative consensus sequences could be identified. Consequentially, only the highest confidence alignments were extracted to produce amino-acid sequences. Protein sequences from the *T. socialis* genome segment, *de novo* transcript and reference-mapped consensus sequence were combined, producing a partial protein of 212 aa in length (relative to ~1000 aa residues in other taxa). The partial *T. socialis* IAR1 protein was aligned with *C. reinhardtii*, *G. pectorale* and *V. carteri* InvA/IAR1 proteins with MAFFT and a 237 aa residue alignment was created where the majority of gaps that corresponded to major gaps in the *T. socialis* sequence were removed. This alignment was used to create a protein phylogeny using a neighbour-joining tree construction method (within CLC Genomics Workbench) with Jukes-Cantor protein distance measurement and with 200 bootstraps performed. It remains uncertain if the *INVA* gene is, in fact, present in *T. socialis*.

9.1.3.3.3 Cyclins

The analysis of cyclins is described in Featherston et al (2018). Dr Yoko Arakaki and Professor Hisayoshi Nozaki improved cyclin gene models and performed the phylogenies presented in Featherston et al (2018). Prior to analysis performed by Dr Arakaki and Professor Nozaki the analysis of cyclins was performed using the genomic data. OrthoMCL, BLAST, InterProScan, HMMER3 and protein phylogenies were used to identify and characterise cyclins (Bisova, et al. 2005; Merchant, et al. 2007; Prochnik, et al. 2010; Hanschen, et al. 2016). Cyclin proteins from *C. reinhardtii*, *G. pectorale* and *V. carteri* were used as queries against the genome of *T. socialis* with tblastn and as queries against the *T. socialis* proteome using blastp ($E=1x10^{-5}$). In addition to BLAST-based homology searches, HMMER3.1 (Finn, et al. 2015) was used to search for un-annotated cyclin D1 genes. A nucleotide sequence alignment of *C. reinhardtii*, *G. pectorale* and *V. carteri* cyclin D1 CDS sequences was generated with MAFFT, which was used creating a cyclin D1 HMM that was scanned against the *T. socialis* genome with HMMER3.1. Model testing for performing protein phylogenies was performed using ProtTest. Phylogenetic trees were drawn for all cyclin proteins identified by homology searches as well as separate trees for D-type cyclins with PhyML (Criscuolo 2011), RAxML, Phylip (Ropelewski, et al. 2010), the neighbour-joining algorithm within CLC Genomics workbench and Mr Bayes.

D-type cyclin proteins were challenging to resolve with phylogenetic methods and therefore a number of phylogenies were drawn that variously used annotations from *C. reinhardtii* v3 and v5.5 as well as for *V. carteri* v1, v2 and v2.1. The RAxML phylogeny was performed with a VT substitution matrix, empirical frequencies, inclusion of invariable sites and gamma distribution with 1000 bootstraps. Putative *T. socialis* D1 cyclins identified from phylogenetic trees were inspected for the presence of LXCXE (retinoblastoma binding motif) and LXXLL (co-activator motif) motifs, proteolytic cleavage sequences rich in proline, glutamic acid, serine and threonine (PEST sequences) as well as cyclin domains (PF00134 and PF02984). Simple motifs (LXCXE and LXXLL) were identified using an awk regulation expression and cyclin domains were identified from HMMER3 scans to the Pfam-A database. The gene structures of volvocine D-type cyclins were examined to characteris *T. socialis* D-type orthologues. WebScipio (http://www.webscipio.org/search) (Hatje, et al. 2013) was used to create gene structures for each *T. socialis*, *C. reinhardtii*, *G. pectorale* and *V. carteri* D-type cyclin gene. The genomes of *C. reinhardtii* (v5) and *V. carteri* (v2) were selected from genomes available on WebScipio while relevant scaffold sequences from *T. socialis* and *G. pectorale* were uploaded to the website. WebScipio employed blat to align provided proteins to genomic sequences and subsequently created gene structures in YAML file format, which map intron and exon boundaries.

A protein multiple sequence alignment of D-type cyclins was generated using MAFFT, which together with YAML gene structure files were submitted to GenePainter (http://www.motorprotein.de/genepainter/) (Muhlhausen, et al. 2015) to map aligned D-type cyclin proteins to their exonic and intronic structures. The gene-structure-informed alignment was subsequently examined for shared and missing exons and introns.

9.1.3.3.4 Retinoblastoma Gene

Methods used to examine volvocine retinoblastoma (*RB*) homologues are described in Chapter 3 (Featherston et al, 2018). Included here are additional analyses that included RB homologues from chlorophytes other than the volvocines.

9.1.3.3.5 Fasciclin domain-containing proteins

Fasciclin domain (PF02469) containing proteins in *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* were identified by querying Pfam-A with HMMER3. Proteins were aligned using MAFFT, model testing was performed with ProtTest and RAxML was used for the generation of a phylogenetic tree. The *V. carteri* algal cell adhesion molecule (Algal-CAM) protein (XP_002958363.1) (Huber and Sumper 1994) was identified in the *V. carteri* v1 proteome and was identified in a protein phylogeny of fasciclin-domain containing proteins. The protein phylogeny identified 3 putative paralogues of *V. carteri* Algal-CAM.

Other fasciclin-domain containing proteins from *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* did not cluster with Algal-CAM paralogues. These results are described in Chapter 3 (Featherston, et al. 2018). To further inspect the evolutionary relationship of these proteins in the volvocine lineage, a preliminary assembly of *P. starrii* was inspected for fasciclin-domain containing proteins. Using HMMER3 a hidden-markov model (HMM) of Algal-CAMs was generated from an alignment of *V. carteri* algal-CAMs (n=3). The HMM of algal-CAMs was used to generate a protein profile for augustus (2.5.5) (Stanke, et al. 2006). Augustus gene modelling was executed against the *P. starrii* genome in protein profile (PPX) mode with the algal-CAM protein HMM profile and *C. reinhardtii* gene model training. Computationally derived *P. starrii* proteins with homology to *V. carteri* algal-CAMs were identified using blastp. Candidate *P. starrii* algal-CAM proteins were included together with fasciclin-domain containing proteins from *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* and a protein phylogeny was generated using the same methods as before.

A single *P. starrii* protein was identified that clustered together with *V. carteri* algal-CAMs. The putative *P. starri* algal-CAM was manually curated for gene structure. A phylogeny of fasciclin-domain containing proteins from *C. reinhardtii*, *T. socialis*, *G. pectorale*, *V. carteri* that included the manually curated *P. starri* putative algal-CAM protein was generated. For the above phylogeny, proteins were aligned using MAFFT

and a phylogeny was generated with RAxML using the Blosum62 model with invariant sites, gamma distribution, and empirical frequencies with 1000 bootstraps.

9.2 Results

9.2.1 Synteny

Previous analyses of synteny in volvocine algae have found that while large regions of conserved synteny can be identified, syntenic rearrangements are extensive. The preservation of synteny between *V. carteri* and *C. reinhardtii* is comparable to synteny between humans and chickens. The fragmented nature of the *T. socialis* genome sequence did not lend itself to a thorough assessment of synteny. Therefore, a limited synteny analysis between *C. reinhardtii* and large (>100kb) *T. socialis* scaffolds was performed. Large syntenic blocks spanning greater than 50kb were observed (Figure 9.1 and Figure 9.2). Synteny between *T. socialis* and *C. reinhardtii* shows a relationship with the size of *T. socialis* scaffolds, which indicates that a more contiguous *T. socialis* is required to more definitively trace syntenic re-arrangements. Another observation is that the rate of syntenic arrangement amongst the volvocines precludes the possibility of performing super-scaffolding based on volvocine synteny.

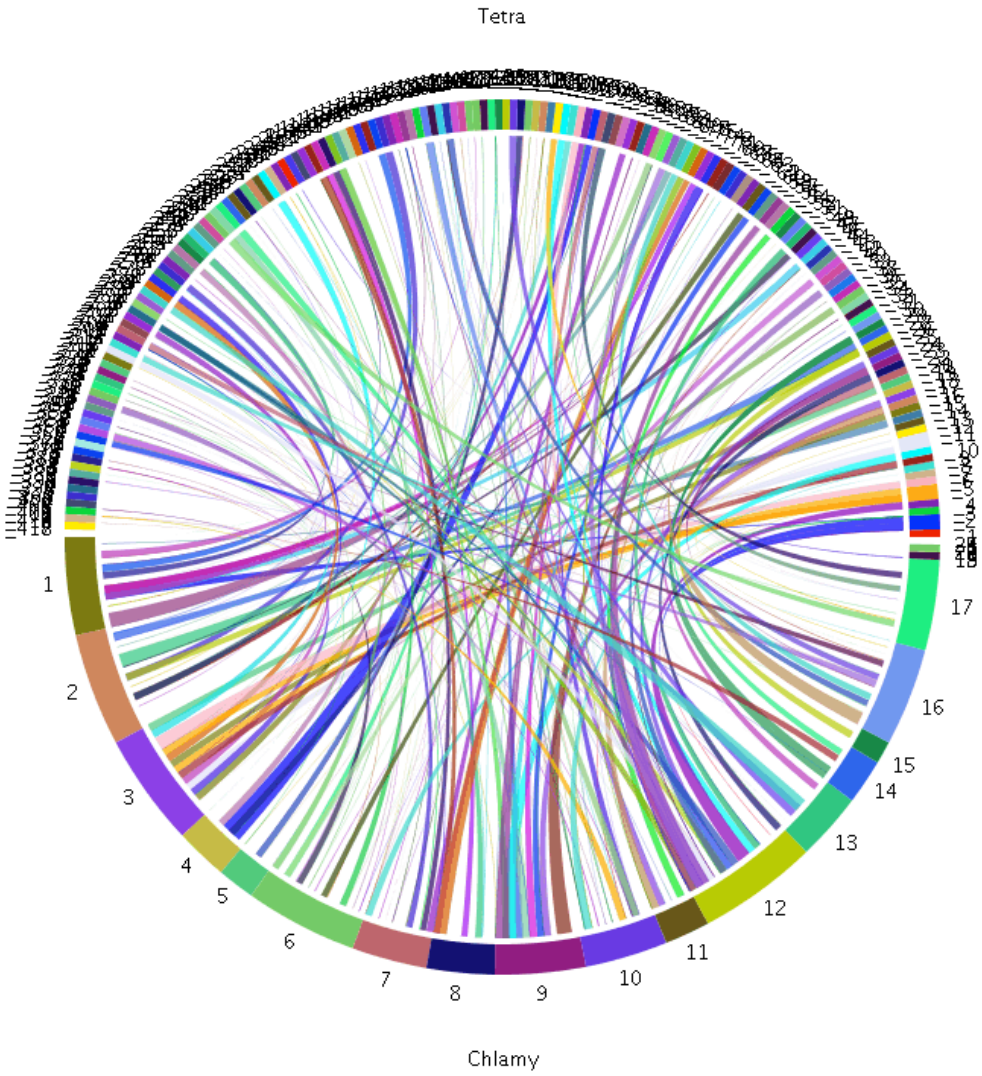


Figure 9.1: Synteny between *Chlamydomonas reinhardtii* and *Tetrabaena socialis* viewed in circular format. Only *Tetrabaena socialis* scaffolds larger than 100kb were used in this comparison.

A number of large regions of syntenic preservation can be found. For example, large regions of synteny between *Chlamydomonas*

reinhardtii chromosome 1 and at least 4 *Tetrabaena socialis* scaffolds can be seen.

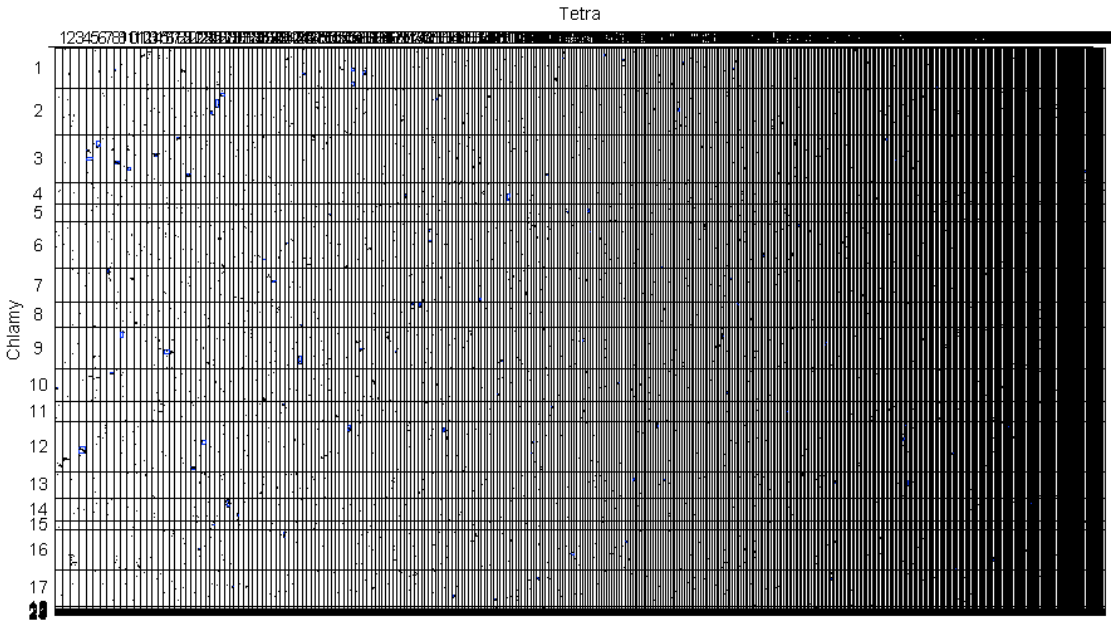


Figure 9.2: Dot plots of synteny between *Chlamydomonas reinhardtii* and *Tetrabaena socialis*. Only *Tetrabaena socialis* scaffolds larger than 100kb were used in this comparison. The dots indicate evidence of synteny. Image generated using Symap.

9.2.2 Phylogenomic analysis of Volvocine Taxonomy

From 18 proteomes OrthoMCL produced 24 949 cluster/families, of which, 192 families were retained after pruning for families containing paralogues and families that did contain proteins from all 18 proteomes i.e. only families with a single orthologue from each taxa were retained (n=192). A RAxML phylogeny generated 100% bootstrap support for all

nodes except for the branch leading to *G. pectorale* and *V. carteri* (53%). The branch leading from *C. reinhardtii* to the multicellular volvocine taxa had 100% bootstrap support (Figure 9.3).

The phylogeny is in agreement with previous phylogenies, which have variously been produced from whole proteome or from conserved genes such as 18S. The phylogeny extends upon the phylogenomic approach used by Hanschen et al (2016) by including both *T. socialis* as well as transcriptome-derived proteomes from additional unicellular taxa belonging to taxa from the chlamydomonadales order. The taxa *P. parva*, *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* form a clade, which agrees with previous findings of these taxa being members of the reinhardtinia-clade (Nakada, et al. 2008). Also apparent is the polyphyletic nature of the genus *Chlamydomonas*.

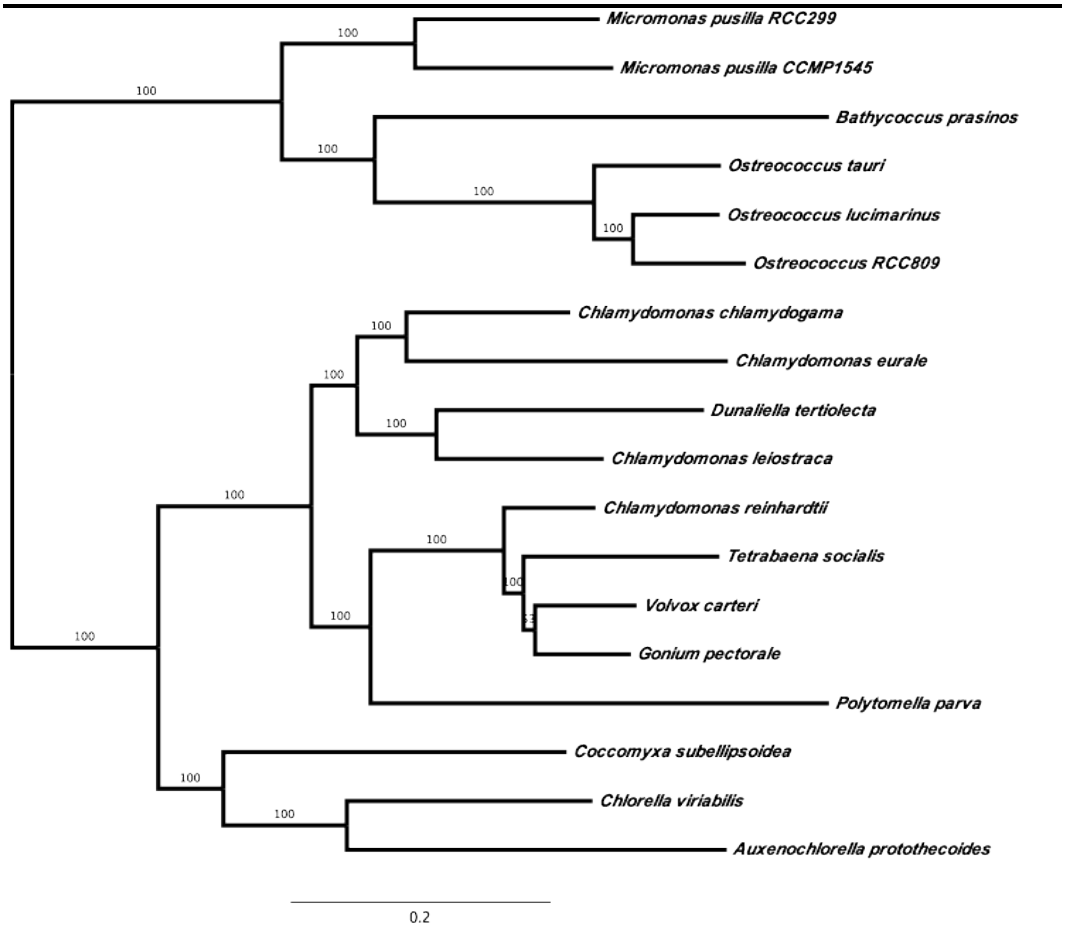


Figure 9.3: RAxML phylogenetic analysis of proteomes from genome sequenced chlorophytes with additional transcriptome derived proteomes from taxa in the Chlamydomonadales order. The phylogeny was generated from 192 strict one-to-one protein orthologues, which totalled 126 722 aligned, trimmed and concatenated amino acid residues.

9.2.3 Comparative Genomics: Global annotations associated with the Evolution of Multicellularity

9.2.3.1 Mercator assigned MapMan annotations

Due to the distance of chlorophytes from plants a relatively low percentage of genes (~30%) were assigned a MapMan functional annotation by Mercator. By comparison, in a previous study 44,6% and 60% of *Solanum pennellii* and *A. thaliana* proteins were respectively assigned MapMan terms (Bolger, Scossa, et al. 2014). The number of proteins with MapMan terms between the examined species was similar (*C. reinhardtii* 30,32%, *T. socialis* 26,8%, *G. pectorale* 30,26% and *V. carteri* 30,28%). Assigned MapMan terms were used for a high-level comparison of annotation differences between the unicellular *C. reinhardtii* and the multicellular volvocines (*T. socialis*, *G. pectorale* and *V. carteri*). A total of 14 functional categories were identified as significantly different between *C. reinhardtii* and the multicellular species while a further 15 and 45 respective MapMan categories were only present in *C. reinha*rdtii or only present in the multicellular species. Amongst significantly different terms only the term “RNA regulation of transcription with TBP” was enriched in the multicellular taxa. Proteins assigned with this term include transducing/WD40 domain containing proteins and *C. reinhardtii* flagellar-related WD related proteins. Domain analysis demonstrated enrichment of WD40 domain containing proteins in multicellular taxa.

The remaining 13 terms were enriched in *C. reinhardtii* of which 2 were related to metabolism (lipid metabolism and flavonoids) and remaining terms are candidates for association with the evolution of multicellularity. A total of 5 terms were related to receptor signalling and kinase activity (serine/threonine kinases and tyrosine kinases). Of the 5 terms involving signalling receptors 4 describe leucine-rich repeat kinases. Prochnik et al (2010) identified that leucine-rich repeat domains are less common in *V. carteri* than in *C. reinhardtii* and Hanschen et al (2016) showed the same to be true for *G. pectorale* while *T. socialis* does not show a decrease (Table 9.1). MapMan annotations indicate that leucine-rich repeat kinases may be related to receptor signalling and, interestingly, the domains are only reduced in the more complex *G. pectorale* and *V. carteri*. Terms enriched in *C. reinhardtii* include the class “calmodulin-related signalling (CRCK1)”. Domain analysis also showed EF-hand 6 and 7 domains to be enriched in *C. reinhardtii*. The MapMan term “Protein degradation cysteine proteases” was identified as enriched in *C. reinhardtii*. Prochnik et al (2010) identified cysteine protease domain-containing proteins as expanded in *V. carteri* relative to *C. reinhardtii*. However, improvements in *C. reinhardtii* annotation (Blaby, et al. 2014) no longer support this and the opposite is now the case such that more cysteine protease domains are identified in *C. reinhardtii* than the multicellular proteomes ((Hanschen, et al. 2016) and (Featherston, et al. 2018)). MapMan annotation highlights a possible role for the cysteine protease domain proteins as part of the protein degradation machinery.

A class of squamosa promotor binding (SBP)-like transcription factors were enriched in *C. reinhardtii* (although SBP-domains were not identified as significantly different). Of the 45 MapMan families unique to the multicellular volvocines only one is shared in all three species, namely "secondary metabolism-sulphur containing", which in *A. thaliana* is involved in FE-S chloroplast assembly.

Table 9.1: MapMan Terms significantly different (p<0.05) between the unicellular *Chlamydomonas reinhardtii* and the Multicellular Volvocines.

MapMan Functional Class Annotation	P-value
Signalling receptor kinases wheat LRK10 like	0.001789
Protein postranslational modification	0.006917
Signalling receptor kinases leucine rich repeat XII	0.007005
Protein postranslational modification kinase receptor like cytoplasmatic kinase VII	0.009665
Stress abiotic unspecified	0.010599
Signalling receptor kinases leucine rich repeat VIII VIII-1	0.010599
Protein degradation cysteine protease	0.022566
Transport phosphate	0.025400
Lipid metabolism FA desaturation desaturase	0.026949
Development squamosa promoter binding like (SPL)	0.028188
RNA regulation of transcription General Transcription, TBP-binding protein	0.029503*
Secondary metabolism flavonoids chalcones	0.037932
Not assigned no ontology proline rich family	0.039003
Signalling receptor kinases leucine rich repeat VIII VIII-2	0.042601

* enriched in multicellular taxa.

9.2.3.2 Gene Ontology Analysis

Gene ontology (GO) terms identified by InterProScan were compared at all levels of GO hierarchy and 9 terms were identified as significantly different ($p<0.05$) between *C. reinhardtii* and the multicellular volvocines (Table 9.2). Noteworthy terms include “membrane” and “transmembrane transport” which were more abundant in *C. reinhardtii* than in the multicellular volvocines. The terms “small GTPase mediated signal transduction”, “peptidase activity”, “hydrolase activity acting on acid anhydrides” and “DNA integration” were significantly enriched in the multicellular species. Both MapMan and GO terms indicate certain signalling processes to be of interest for evolution of multicellularity in the volvocine lineage.

Table 9.2: Gene Ontology terms significantly different (p<0.05) between the Unicellular *Chlamydomonas reinhardtii* and the Multicellular Volvocines.

GO term	Term Label	P-value
GO:0016817	hydrolase activity, acting on acid anhydrides	3.19E-04
GO:0015074	DNA integration	1.07E-03
GO:0004114	3',5'-cyclic-nucleotide phosphodiesterase activity	1.89E-03
GO:0007264	small GTPase mediated signal transduction	1.83E-02
GO:0008233	peptidase activity	2.44E-02
GO:0016020	membrane	2.90E-02
GO:0055085	transmembrane transport	3.58E-02
GO:0008484	sulfuric ester hydrolase activity	4.23E-02
GO:0008061	chitin binding	4.37E-02

9.2.3.3 Specific Gene Families

The Prochnik (2010) genes "involved in processes that are associated with increased developmental complexity in *V. carteri* relative to *C. reinhardtii*" (hereafter Prochnik genes) were identified for *T. socialis*. Pherophorins, Volvox Metalloproteases (VMPs) and cell cycle genes are described separately. Besides pherophorins and VMPs approximately 138 orthologues were identified from the predicted proteome of *T. socialis*. A further 4 genes were identified from TBLASTN searches against the *T. socialis* genome assembly that were not identified by automated gene prediction using MAKER2.

Only 2 genes (*CDKE1* and *CDKH1*) that might be expected to be present in *T. socialis* could not be identified in the assembly. Additional paralogues for a few Prochnik genes were identified in *T. socialis* that are either the result of species-specific gene expansions or perhaps have remained un-annotated in other taxa (Featherston, et al. 2018). Some noteworthy observations identified from the analysis of Prochnik protein orthologues include:

Protein secretion and membrane trafficking:

- *rabG1* and *rab23* (small Rab-related GTPase) are absent from *C. reinhardtii* and present in all multicellular taxa.
- *VAMP74* (R-SNARE protein) is present in *C. reinhardtii* and absent from all multicellular taxa.

Actin cytoskeleton:

- *myoD* (myosin D) is present in all multicellular taxa and absent in *C. reinhardtii*.
- *myoF* is only present in *V. carteri*.
- *nap1* (novel actin-like protein) is present in all multicellular taxa and absent in *C. reinhardtii*.
- *arp6* (Actin-related protein) is absent in *C. reinhardtii* and *T. socialis* and present in *G. pectorale* and *V. carteri*.

Microtubule cytoskeleton:

- Four copies of *TUA* (alpha tubulin A) are present in *T. socialis* whereas two copies are present in other taxa.
- *map1* (microtubule associated 1) is absent in *C. reinhardtii* and *T. socialis* and present in *G. pectorale* and *V. carteri*.
- Four copies of *DHC1B* (cytoplasmic dynein heavy chain 1b) are present in *T. socialis* whereas 1 copy is present in other taxa. However, additional dynein heavy chain proteins have been identified in volvocine proteomes and an OrthoMCL cluster of dynein heavy chain proteins includes 14 *C. reinhardtii*, 23 *T. socialis*, 27 *G. pectorale* and 6 *V. carteri* proteins.
- *ttl4* (tubulin-tyrosine ligase) is present in all multicellular taxa and absent in *C. reinhardtii*.

Basal body proteins

- *BBS7* (Bardett-Beidl syndrome 7) is present in *C. reinhardtii* and *T. socialis* and absent in *G. pectorale* and *V. carteri*.1

Kinesins

- The *kif6*-type kinesis-like protein is present in all multicellular taxa and absent from *C. reinhardtii*.
- A putative *kif9*-type kinesin is only present in *V. carteri*.
- 3 unclassified kinesin proteins, an Atk5 related kinesin and a Kar3 member kinesin-like protein are found only in the multicellular taxa.

- The Kif4A protein is absent in *C. reinhardtii* and *T. socialis* and present in *G. pectorale* and *V. carteri*.
- A possible IARA protein was partially recovered for *T. socialis.* Nevertheless, the protein phylogeny of INVA correlated with the phylogenetic position of the four taxa (*C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri* – Figure 9.4). It is unknown if the protein is intact in *T. socialis*, however, RNAseq data indicate the transcript is expressed.

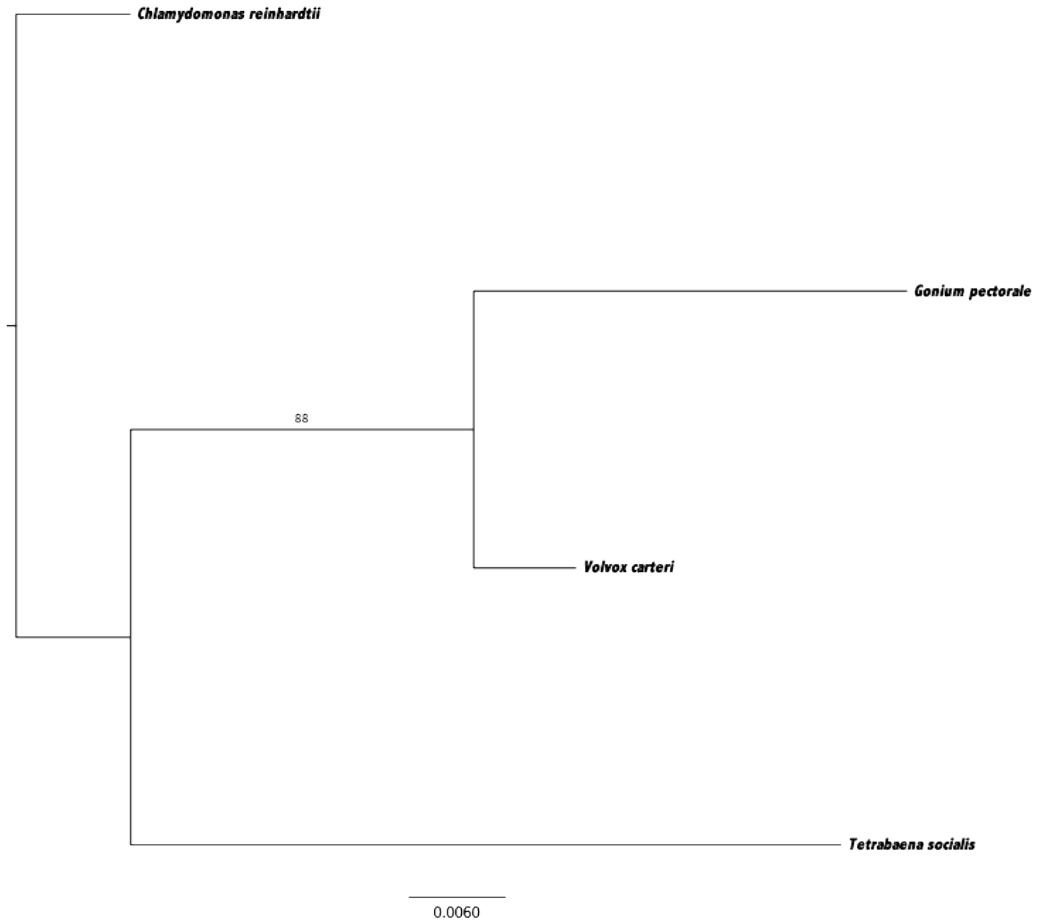


Figure 9.4: Neighbour-joining phylogeny of InvA/IAR1 produced from an alignment of 237 amino acids.

Cell wall proteins

- The Gp1 cell wall protein is absent in all multicellular taxa.
- The Gp3 cell wall protein is present in all taxa.

Initial analyses proved challenging to resolve D1-type cyclins (Figure 9.5) but are described in Featherston et al (2018). Genome guided sequencing of D-type cyclins aided in generating improved phylogenies. Even so, similar to Hanschen et al (2016) low bootsrap support (<50%) was obtained for the branch leading to D1-type cyclins. Examining shared exons and introns were used to aid in identifying expanded D1-type cyclins in *G. pectorale*. The same methodology was followed as per Hanschen et al (2016) but with the inclusion of *T. socialis* genes. A number of proteins from *T. socialis*, *G. pectorale* and *V. carteri* were not included in the analysis due to incomplete gene models. However, D-type cyclins were included from most taxa. With the exception of CYCD3, the alignment of proteins together with their corresponding gene structures did not identify exons or introns unique to a particular D-type cyclin subtype.

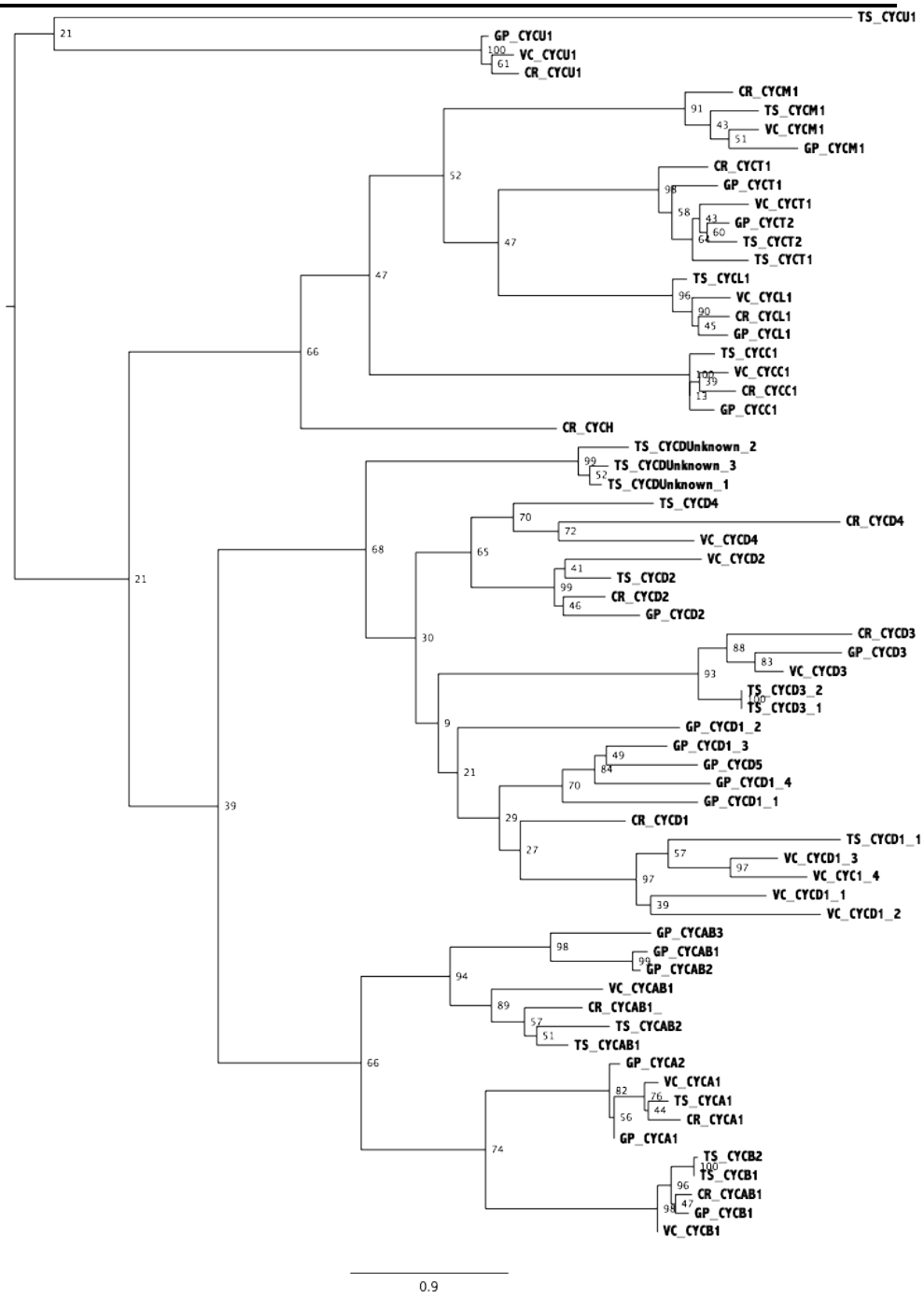


Figure 9.5: An example RAxML maximum likeilhood phylogeny of volvocine cyclins. Tetrabaena socialis cyclins that were as yet not positively identified as D1-type cyclins are labelled as unknown.

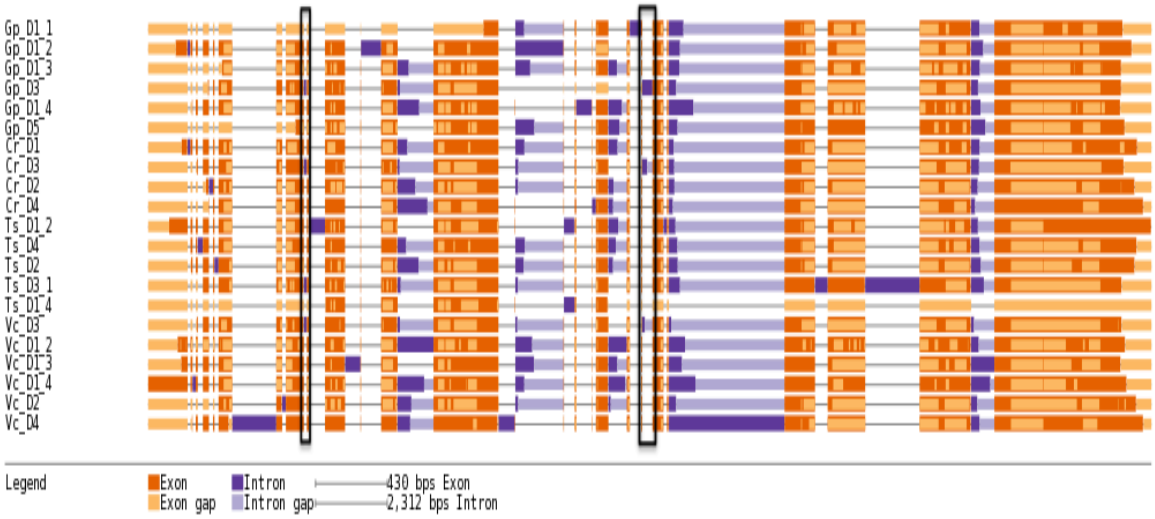


Figure 9.6: Gene structures of various cyclin D proteins. Intron positions unique to cyclin D1 genes were not identified. However, unique introns are present in cyclin D3 genes and demarcated with black brackets.

9.2.3.4 The Retinoblastoma Homologue

A MAFFT alignment of MAT3/RB proteins revealed that as per Hanschen et al (2014) when compared with *C. reinhardtii* and *C. incerta* that colonial/multicellular taxa possess a shortened linker region between the RB-A and RB-B domains. The linker region in *Y. unicocca* is similar to the linker region in *C. reinhardtii* –in terms of both the composition of amino acid (aa) residues and length (106aa and 130aa respectively). The linker region of the more distantly related unicellular *Chlamydomonas chlamydogama* was 62aa in length.

The length of linker regions in volvocine MAT3 regions is presented in table 9.3. In addition to MAT/RB proteins generated by Hiraide et al (2013) orthologues of MAT3/RB protein were identified from other unicellular chlorophytes, namely: *Bathycoccus prasinos*, *Ostreococcus lucimarinus*, *Micromonas pusilla* (strains 299 and 1545), *Chlamydomonas chlamydogama* and *Coccomyxa subellipsoidea*. The length of the linker region in these more distantly related taxa varied from 67 to 129aa. These additional orthologous were not included in Featherston et al (2018). Phosphorylation targets within the linker region between domains B and C do not correlate with the evolution of multicellularity (Table 9.4)

Table 9.3: The length of the linker region between the retinoblastoma A and B domain in Chlorophyte orthologues of MAT3/RB. A (+) or (-) symbol next to the species name indicates mating-type plus or minus.

Species	Length of Linker Region
***Chlamydomonas reinhardtii* +**	130
***Chlamydomonas reinhardtii* -**	130
Chlamydomonas incerta	134
***Yamagishiella unicocca* +**	106
***Yamagishiella unicocca* -**	106
***Tetrabaena socialis* (homothallic)**	76
***Eudorina spp* Female**	93
***Eudorina spp* Male**	93
***Volvox Africana* Male**	93
***Volvox Africana* Female**	93
***Pleodorina starrii* Female**	96
***Pleodorina starrii* Male**	96
***Volvox carteri f. nagariensis* Male**	83
***Volvox carteri f. nagariensis* Female**	47
***Volvox carteri f. weismannia* Male**	81
Volvox carteri f. kawasakiensis Male	88
***G. pectorale* +**	47
***G. pectorale* -**	47
***V. observus* Male**	87
Chlamydomonas chlamydogama	62
Bathycoccus prasinos	129
Ostreococcus lucimarinus	70
Ostreococcus lucimarinus RCC809	67
Micromonas puscilla (299)	91
Micromonas puscilla (1545)	86
Coccomyxa subellipsoidea	108

Table 9.4: Phosphorylation targets within the linker between the retinoblastoma B and C-terminal domain in Chlorophyte orthologues of MAT3/RB

Species	Typical phosphorylation targets (M and F indicate male and female)	Atypical phosphorylation targets (M and F indicate male and female)
Chlamydomonas reinhardtii	SPTK (+/-)	SPRG (+/-), TPTG (+/-)
Chlmaydomonas globator	SPTK (+/-)	SPRG (+/-)
Bathycoccus prasinos	TPKR, TPNR, SPIR, SPMR	
Ostreococcus lucimarinus	SPMR, SPRR	TPRT
Micromonas pusilla 299	SPAR	TPRS, SPMR, SPSG, SPLP, SPAS
Micromonas pusilla 1545	SPRK	SPAQ, TPGS, TPLP, SPSP, SPMR, SPPA
Coccomyxa subellipsoidea	SPSK, SPMR, TPLR	SPAL, TPGG, SPSQ, TPPA, TPLG
Chlamydomonas chlamydogama	SPAK	
Tetrabaena socialis	SPLK, SPLR	SPRG
Gonium pectorale	SPSK (+/-)	
Yamagishiella unicocca		SPRA (+/-)
Eudorina spp		SPWA (M/F), TPLL (M/F), SPQG (M/F)
Volvox africanus	SPVR (M/F)	SPVL (M/F)
Volvox observus	None identified (protein incomplete)	
Pleodorina starrii		SPNG (M or F)
Volvox carteri f. nagariensis	TPGK (F)	SPDN (F), SPGT (F)
Volvox carteri f. weismannia		SPRG (M)
Volvox carteri f. kawasakiensis		SPTA (M&F), SPRG (M&F)

9.2.3.5 Fasciclin-domain Containing Proteins

Gene family birth and death analysis in Featherston et al, (2018) identified a family of fasciclin-domain contain proteins that originated at the ORIMC. To further examine fasciclin-domain containing proteins in the lineage, fasciclin-domain containing proteins were identified from *C. reinhardtii* (N=21), *T. socialis* (N=18), *G. pectorale* (N=17) and *V. carteri* (N=17) and an RAxML phylogeny was generated. The phylogeny identified that 2 additional proteins (V127393 and V108097) cluster together with the annotated *V. carteri* algal-CAM protein (V127389) in the absence of other proteins. The family identified by orthoMCL and Count as originating at the ORIMC contains two *T. socialis* proteins (TSOC_001029-RA and TSOC_001027-RA) and one protein from both *G. pectorale* (scaffold00001.g603t.1) and *V. carteri* (96911). Using the closest *C. reinhardtii* protein (Cre17.g728350.t1.1) as a reference for the gained family, codeml identified significant ($p<0.05$) evidence for codon-level positive selection. Previous findings suggest a role for the algal-CAM protein(s) in facilitating embryo inversion in *V. carteri* and thus the draft genome of *P. starrii* (which undergoes inversion) was inspected for orthologues of algal-CAM. A candidate protein was identified in *P. starrii*, the protein was manually curated for structure and was included in another phylogeny. The phylogeny (Figure 9.7) shows that the *P. starrii* protein clusters with *V. carteri* algal-CAMs and may be orthologous.

The alignment of fasciclin-domain proteins shows that the protein sequence from position 109-151 of *V. carteri* paralogues and from position 90-133 of the *P. starrii* protein is distinct from other fasciclin-domain proteins.

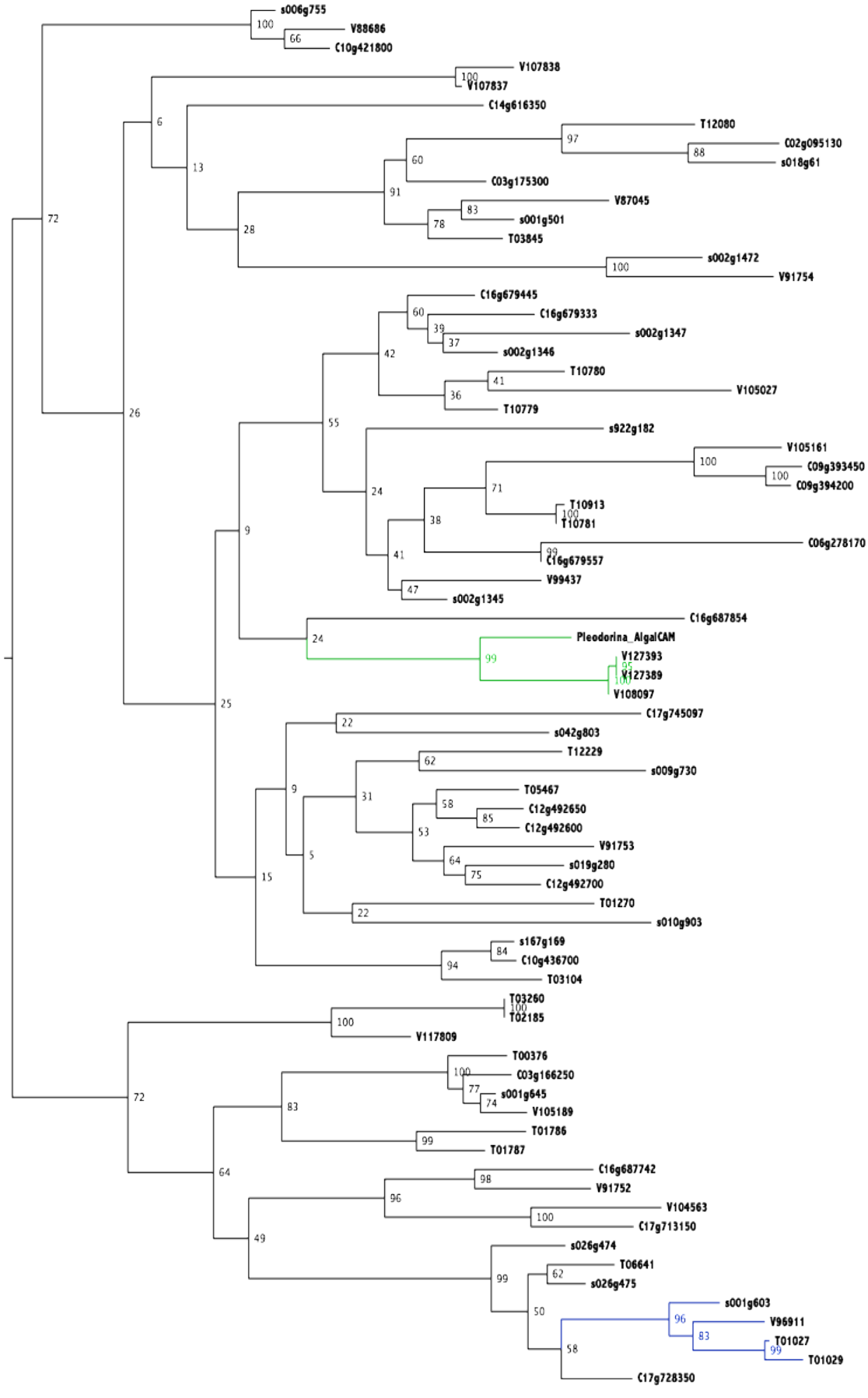

s006g755
V88686
C10g421800
V107838
V107837
C14g616350
T12080
C02g095130
s018g61
C03g175300
V87045
s001g501
T03845
s002g1472
V91754
C16g679445
C16g679333
s002g1347
s002g1346
T10780
V105027
T10779
s922g182
V105161
C09g393450
C09g394200
T10913
T10781
C06g278170
C16g679557
V99437
s002g1345
C16g687854
Pleodorina_AlgalCAM
V127393
V127389
V108097
C17g745097
s042g803
T12229
s009g730
T05467
C12g492650
C12g492600
V91753
s019g280
C12g492700
T01270
s010g903
s167g169
C10g436700
T03104
T03260
T02185
V117809
T00376
C03g166250
s001g645
V105189
T01786
T01787
C16g687742
V91752
V104563
C17g713150
s026g474
T06641
s026g475
s001g603
V96911
T01027
T01029
C17g728350
0.4

Figure 9.7: Mid-point rooted phylogeny of fasciclin-domain containing volvocine proteins. Included in the phylogeny are proteins from *Chlamydomonas reinhardtii* (labelled CXXgXXXXXX), *Tetrabaena socialis* (labelled TXXXXX), *Gonium pectorale* (labelled sXXgX*), *Volvox carteri* (labelled VXXXX*) and a single protein from *Pleodorina starrii* (labelled as Pleodorina_AlgalCAM). Labelled in blue are is a branch that contains an OrthoMCL family of fasciclin-domain containing proteins identified by Count as originating at the origin of multicellularity (ORIMC). A *Chlamydomonas reinhardtii* homologue (Cre17.g728350) is immediately adjacent to the novel ORIMC family. Green branches include *Volvox carteri* algal-CAM paralogues as well as a putative algal-CAM identified in *Pleodorina starrii*.

9.3 Discussion

9.3.1 Syntenic rearrangements

Syntenic re-arrangements could not be adequately assessed for *T. socialis* due to the draft nature of the genome sequence. However, all analyses of synteny in the lineage (Prochnik, et al. 2010; Hanschen, et al. 2016) suggest that extensive syntenic re-arrangements have occurred and these may have important evolutionary implications (Heng 2009).

9.3.2 Phylogeny

The phylogenetic tree produced in the current analysis is consistent with and corroborates the findings of phylogenetic analyses derived from smaller numbers of genes as well the phylogenomic analyses by Hanschen et al (2016). In principle, due to the greater sequence space a phylogenomic analysis is capable of creating a phylogeny of greater resolution and accuracy than more conventional approaches that utilise a limited number of phylogenetically informative genes. Phylogenetic analyses of the Chlamydomonadales order has largely centred on the use of phylogenetically informative genes (e.g. *rbcl*, *cox1*, *18S* etc) although Hanschen et al (2016) produced a phylogenomically-derived phylogeny from 11 taxa (3 from the Chlamydomonadales order) that was consistent with previous findings. By including the proteome of *T. socialis* and by including transcriptome derived proteomes from 5 additional unicellular Chlamydomonadales species the current analysis extends upon the analysis by Hanschen et al (2016). These inclusions provide further resolution of the Chlamydomonadales order. The phylogeny confirms; (i) the relatedness of *C. reinhardtii* to the multicellular volvocines, (ii) that the genus *Chlamydomonas* is polyphyletic and (iii), that it is unlikely that *C. reinhardtii* is a unicellular derived form of *G. pectorale*. The inclusion of the *Polytomella* genus with *C. reinhardtii* and the multicellular volvocines in a clade known as the reinhardtinia-clade is confirmed in this analysis (Nakada, et al. 2008).

9.3.3 High Level Descriptors

Comparisons of functional annotations were not expected to identify specific molecular mechanisms associated with evolution of multicellularity. Broad comparisons between the unicellular *C. reinhardtii* and the multicellular taxa of Mercator terms and GO terms did however provide “high-level” insights into the evolution of multicellularity in the volvocine lineage. Comparisons were explicitly limited to direct comparisons between *C. reinhardtii* and the multicellular volvocines as the evolution of multicellularity was considered lineage specific rather than a ‘general program for multicellularity’ that might apply across all chlorophytes. Assignment of annotations is predominantly based on homology and is itself a challenging process and is likely to be error prone; including both type I and type II errors (false positives and false negatives) regardless of the approach used. Nevertheless, broad comparisons in the volvocines identified a few relevant observations. Differentially represented MapMan and GO terms between *C. reinhardtii* and the multicellular volvocines include a number of terms related to signal transduction and receptor signalling (e.g. GO:0007264 and 5 mercator terms). Although domain content is similar between *C. reinhardtii* and the multicellular taxa differences do point towards signalling pathways being modified.

9.3.4 Fasciclin domain containing proteins

Previous work in *Volvox* identified a cell adhesion molecule – algal cell adhesion molecule 1 (CAM1) - that affects cell adhesion during embryogenesis and embryo inversion. Fasciclin-domain containing proteins are also constituents of the flagellar proteome of *C. reinhardtii*. Phylogenetic analysis of fasciclin-domain containing volvocine proteins revealed the family gained at the ORIMC is related to *C. reinhardtii* homologues but not closely related to *C. reinhardtii* flagellar proteins and is closer to known *V. carteri* cell adhesion molecules. Recent work by Marriage et al (2015) revealed that when a specific *G. pectorale* fasciclin-domain gene was transformed into *C. reinhardtii* simple (as yet undescribed) colonial forms of *C. reinhardtii* mutants arise (Marriage and Olson 2015). The initial formation of simple colonially multicellular forms is predicated on the increased propensity for cell adhesion post embryogenesis and cell adhesion molecules may be an important molecular component of this phenotype. Furthermore, phylogenetic analysis of fasciclin-domain containing proteins demonstrated that the *Volvox* CAM1 protein forms a distinct cluster with two other *V. carteri* proteins that do not cluster with proteins in *C. reinhardtii*, *T. socialis* and *G. pectorale*. *Volvox* CAM1 has been demonstrated to be important for cell adhesion after division and embryo inversion. Embryo inversion is a trait absent from other volvocines algae with sequenced genomes and in the current study the phylogeny of fasciclin domain-containing proteins shows that the *Volvox* CAM1 protein/s are divergent from other taxa.

Volvox CAMs are therefore good targets for further investigation of additional molecular constituents required for embryo inversion and could play an important in this process. As the genomes of other volvocine algae become available it will be intriguing to explore the role of algal-CAMs in relation to embryo inversion. In particular, if the *V. carteri* CAM1 proteins are shared by other members of the Volvocaceae, including from taxa such as *Volvox globator*, which undergoes type II embryo inversion.

10. APPENDIX D: GENOMIC INSIGHTS OF KIRK'S 12 STEPS

Genome sequences alone are insufficient for explaining specific developments in organismal complexity because the genotype–phenotype correlation is a direct causal relaitonship. However, genome sequences are useful for identifying genes that correlate or associate with specific organismal developments. Identifying genes that correlate with specific volvocine developmental traits is the aim of volvocine genome research, however, this remains challenging because even the arguably most complex organism, *V. carteri*, is highly similar at the proteomic level to the unicellular *C. reinhardtii*. Despite the limitations of genome-level research additional genomes from volvocine algae can provide greater resolution into the nature and timing of genomic events that underpin volvocine innovations. With this in mind, the current study added the genome of *T. socialis* to aid in the understanding of early volvocine traits. By contrasting the genomic repertoire of simple colonial organisms with the *V. carteri* repertoire greater insights in later development can be obtained. What follows is an attempt to correlate genomic developments with the 12 steps in developmental complexity outlined by Kirk (Kirk 2005).

10.1 Incomplete Cytokinesis

The most important aspects of incomplete cytokinesis are the delay of cleavage after cell division and for cell-to-cell adhesion molecules to

retain cellular connection. Neither of these are fully understood. Incomplete cytokinesis may occasionally occur in *C. reinhardtii* (Kirk 2005). In the colonial organisms cleavage is delayed until the following cycle of cell division and this modification is likely to involve an alteration in the cell cycle program. The cell cycle aspects of the volvocine algae will be discussed below under the header 'Genetic control of cell number'. From the simple 4-celled *T. socialis* to *V. carteri* cytoplasmic bridges are an integral aspect of volvocine cellular connectivity. In *V. carteri*, these bridges are lost in the mature embryo and the ECM is solely responsible for maintaining organismal integrity. However, in the simpler taxa and other species of *Volvox* cytoplasmic bridges are in most instances retained in the mature form. While the composition of volvocine cytoplasmic bridges is not fully known, ultrastructural analysis of developing *Volvox* embryos show that the formation of cytoplasmic bridges is due to incomplete cytokinesis and that the most likely constituents are vesicles that extend from cell to cell (Kirk 2005). In addition, microtubules may also traverse along bridges and in so doing contribute towards cell adhesion. As yet, little is known about how incomplete cytokinesis is genetically controlled.

Microtubule genes of multicellular taxa match almost exactly those of *C. reinhardtii*. However, the alpha tubulins, for which 2 duplicates (alpha tubulin 1 and alpha tubulin 2) are found in *C. reinhardtii*, *G. pectorale* and *V. carteri*, are present as 4 copies in *T. socialis*. The protein sequence of

volvocine alpha tubulins is highly conserved, even in the duplicated *T. socialis* copies. If the observed duplications are important in *T. socialis* they most probably play a role in dose response and are unique to *T. socialis*. Besides cytoplasmic bridges cell adhesion molecules (CAMs) may be important for linking volvocine cells together. Cell adhesion molecules are important in other lineages (Niklas 2014) and a *V. carteri* fasciclin domain-containing algal-CAM has been identified. Antibodies targeting *V. carteri* algal-CAMs applied prior to embryogenesis was shown to severely disrupt embryo morphology. The algal-CAMs were identified to localise to cell-cell interfaces. Huber and Sumper (1995) considered the *V. carteri* algal-CAM to be membrane associated by either protein-protein interactions or by protein-carbohydrate interactions. In addition, Marriage et al (2015) demonstrated that transformation of *C. reinhardtii* with a particular *G. pectorale* fasciclin protein resulted in incomplete cytokinesis and colony formation. The identity of the *G. pectorale* fasciclin gene has not been made publically available, however, Marriage did reveal the identity of this gene (personal communication).

The gene is closely related to another *C. reinhardtii* gene without a clear orthologue in *T. socialis* and *V. carteri* - possibly indicating that it is important for the simpler colonial forms but not in *V. carteri* where ECM deposition may replace its role in cell adhesion. The current study demonstrated that the *V. carteri* CAMs that were targeted by Huber and Sumper's antibodies are unique to *Volvox* (or other Volvocaceae) and

therefore unlikely to be significant in the initial development of incomplete cytokinesis. However, gene family birth and death analysis in the current study identified a cluster of fasciclin domain containing proteins that are unique to the multicellular taxa. Further phylogenetic examination revealed a close relationship of the multicellular proteins to an as yet un-annotated *C. reinhardtii* protein (Cre17.g728350). The relationship to a *C. reinhardtii* homologue indicates that the "gained family" is not novel, however, the multicellular orthologues are sufficiently diverged from the *C. reinhardtii* homologue as to not cluster with MCL clustering. Furthermore, bootstrap support also indicates that the multicellular versions of this protein are more closely related to one another but more distantly related to the *C. reinhardtii* homologue. When including the *C. reinhardtii* orthologue there is also codon-level evidence for positive selection. Although these results are not definitive for a role played by the "gained family" in cell adhesion and as a molecular explanation for incomplete cytokinesis, the totality of evidence does suggest a role for fasciclin proteins in volvocine cell adhesion.

It is also possible that mechanisms important for the formation of simple colonial forms are distinct from those involved in the more complex *V. carteri*. After all, the morphological and embryological characteristics of these taxa are distinct. Additional factors related to incomplete cytokinesis were not identified in the current study.

10.2 Partial and Complete Inversion

Tetrabaena socialis is the first multicellular volvocine to be sequenced that is completely devoid of inversion. With the addition of the *T. socialis* genome there are now genetic resources for a unicellular organism, a multicellular organism devoid of inversion, a multicellular organism with partial inversion (*G. pectorale*) and a multicellular organism with full inversion (*V. carteri*). Hence, such comparative genomics may be exploited to identify additional genes involved in inversion. The role of the kinesin invA in volvocine inversion has been identified from the screening of *V. carteri* inversion-less mutants. Two additional mutants; invB and invC have also been identified. These mutants present different morphological characteristics than invA mutants and are considered to be important for the enlargement of the innermost layer of the ECM, which is required to create sufficient space for inversion to take place, whereas the INVA protein localises to microtubules within cytoplasmic bridges and is considered to actively generate torsion of the said bridges that force cells to rotate around the bridges.

As such, *invA* may be considered to contribute more directly towards the inversion process. The *invA* gene is another example in the volvocines for co-option of a gene with an ancestral function to perform a novel function. An *invA* orthologue, *IAR1*, is present *C. reinhardtii* and can even restore wild-type inversion *V. carteri invA* mutants. To date, the function of *C. reinhardtii IAR1* as well the evolutionary history of the gene

has not been known. The gene was identified in *G. pector*ale and also in the current study in *T. socialis*. Furthermore, in the current investigation putative *IAR1*/*invA* orthologues were identified in the transcriptomes of all available members from the Chlamydomondales order – *Chlamydomonas leiostraca*, *Chlamydomonas chlamydogama*, *Chlamydomonas eurale*, *Dunaliella tertiolecta* and *Polytomella parva*. Orthologues were not identified in *Ostreococcus*, *Micromonas* spp., *Chlorella* spp. or in the prasinophyte *Bathychlamys*. Although the function of *IAR1* in *C. reinhardtii* is only speculative the taxonomic distribution of this gene indicates that its function is limited to members of the Chlamydomondales order and its presence in all assayed taxa of this order is indicative of a highly conserved function spanning hundreds of millions of years (the age for the formation of the Chlamydomondales order is not known but the last common ancestor of *Chlorella* and *Chlamydomonas spp*. may be as distant as 800mya).

Besides genes identified from mutation screens such as *invA* algal-CAMS play a role in inversion. When an antibody targeting *V. carteri* algal-CAM was applied after the 4-celled stage of embryogenesis in a synchronised culture inversion was impeded and the resulting embryos were “halfway inverted”. Otherwise the formation of reproductive cells and other aspects of *V. carteri* morphology were unaffected. Three highly similar algal-CAM molecules were identified in *V. carteri* and, in addition, a single *P. starrii* orthologue was identified. As already mentioned,

orthologues of the *V. carteri* algal-CAMs are not found in *T. socialis* and *G. pectorale*. While a *C. reinhardtii* protein was identified to cluster nearby to *V. carteri* algal-CAM molecules (with 24% bootstrap support) alignments suggests that this protein is not an orthologue.

Although the *P. starrii* gene model is incomplete it demonstrates a close relationship to the *V. carteri* algal-CAMs (one of which is likely derived from an incomplete model), which shows that amongst sampled taxa this form of fasciclin-domain gene is only present in the Volvocaceae and as such are likely to be evolutionarily novel genes. There are many developmental features unique to the Volvocaceae family and hence a gene or gene family unique to the family could be associated with any of these features or indeed may be unrelated to developmental processes. However, inversion is a key aspect of almost all members of this family (except for *Platydorina*) and the work by Huber and Sumper identified a role for algal-CAMs in inversion. Therefore, the association of the algal-CAM genes with the Volvocaceae is highly suggestive that one of their main functions is integral to inversion.

It will be interesting to see if they are absent from *A. gubernaculifera* and *Platydorina* and if they are present in other Volvocaceae such as species *Volvox* (e.g. *Volvox globator*) that undergo type B inversion (Hohn and Hallmann 2011). The role of these proteins in inversion is less well known and Huber and Sumper did not suggest how

or why these molecules were important for inversion. Presumably for inversion to proceed normally cell connectivity must be severed, with only cytoplasmic bridges maintaining connections. So what role could cell adhesion molecules play in this process? A possible scenario is that cell connectivity via CAMs is weak but important and is severed during inversion by the torsion InvA creates. After severing of CAM connections cells invert relative to one another, however, without CAMs cells are not "sticky" and cannot reconnect and hence the phenotype of a halfway inversion results. Furthermore, global comparative genomic analyses as well as analysis of Prochnik genes identified the absence of the *kif4-A* gene in *C. reinhardtii* and in *T. socialis*. The kinesin motor protein encoded by this gene is another candidate to examine in relation to inversion - including partial inversion.

10.3 Rotation of Basal Bodies

This morphological development involves the rotation of basal bodies that re-position flagella from the unicellular anti-parallel position into a parallel position that alters the swimming pattern from a breast-stroke motion in unicellular taxa into a parallel positioning required for coordinated swimming in multicellular taxa (Coleman 2012). Little is known about the molecular mechanisms that control basal body positioning and indeed it is unknown whether basal body re-orientation is genetically programmed. The flagella proteome of *C. reinhardtii* has been determined (Pazour, et al. 2005) but the protein content of the flagella cannot alone

elucidate how the flagella (or spoke head) or the basal bodies that anchor the flagella assemble and therefore how their assembly may be altered. Genes important in flagella formation in *C. reinhardtii* include the *UNI2*, *UNI2* and *UNI3* genes (Coleman 2012). Mutations in these genes are known to result in a uni-flagellate phenotype. To date, the *UNI1* gene/protein has not been published and no consistent differences between the *UNI2* and *UNI3* genes of *C. reinhardtii* and the multicellular taxa have been identified. The current study has not been able to further our understanding of how basal bodies re-orientate in the multicellular taxa. It is possible that that differences in how these basal body and flagella proteins are expressed is the deciding factor in how these proteins re-orientate.

It can be said of *T. socialis* and *G. pectorale* that for both organisms rotation of basal bodies is partial. In the case of *T. socialis* the partial nature of re-orientation is because the angle is not completely 90 degrees (Arakaki, et al. 2013) as in more advanced taxa and in *G. pectorale* the fact that only certain cells show the re-orientation is an alternative case of partial re-orientation (Iida, et al. 2013). In this regard *G. pectorale* is indicative of how this occurs because the underlying genetic mechanisms for both parallel and anti-parallel positioning is encoded. This hints that rather than differences in genomic coding portential that regulatory or cellular morphological factors might explain differences in basal body rotation. Possibly, the varying nature of the volvocine cell walls explains

the difference in how basal bodies orientate. Remarkably, when colonial volvocine zoospores/gametes are liberated of their cell walls during germination basal bodies re-orientate to the anti-parallel configuration and swimming is in a breast stroke style (Coleman 2012). This points towards the cell wall being an important factor in how basal bodies orientate and points towards the process of basal body re-orientation not being an active programmed process. It is not clear that any differences in the cell wall of interior versus anterior *G. pectorale* cells exist, however, the anterior cells are somewhat more round in shape when compared to the inner cells that are more droplet shaped. If cell wall modifications such as the loss of the outer cell wall are also critical for basal body re-orientation could a single genomic development have resulted in both traits? Both traits are equally critical for the functioning of a colonial volvocine and may have evolved together.

10.4 Establishment of Organismal Polarity

Organismal polarity serves different roles in the lineage. In *T. socialis* the simplest form of organismal polarity is present – flagella-ends point outwards instead of inwards (Arakaki, et al. 2013). In *G. pectorale* organismal polarity is found as a difference in flagella orientation between in the interior and anterior cells (Iida, et al. 2013). In the more complex spherical forms with more than 32 cells reproductive cells are more common in the posterior while motile soma are positioned more towards the anterior (Kirk 2005). Such arrangements are critical for motility and the

arrangement in multicellular taxa forces the beating of flagella towards the posterior - driving the organism forward. In contrast with other spherical volvocines in *A. gubernaculifera* the posterior pole houses 4 somatic cells, which suggests a different mechanism for organismal polarity in this organism (Yamashita, et al. 2016). Clearly, a number of mechanisms may be at play throughout the lineage, which makes the identification of genetic causes challenging. In this study, no molecular aspects of this development were identified. However, some thoughts on the matter as well on the general formation of the spherical shape are discussed. Anterior-posterior polarity is established during embryogenesis. From morphological and ultastructural descriptions of embryogenesis in simpler taxa such *T. socialis*, *G. pectorale*, *A. gubernaculifera* and *Eudorina*, it would appear that the shape of cells after division largely affects their position (Arakaki, et al. 2013; Yamashita, et al. 2016).

Generally, for the majority of volvocine taxa, cells are not perfectly round and a cell level anterior to posterior asymmetry exists that results in a droplet/elliptical shape with the narrow anterior apical end facing inwards and the broader posterior bulging outwards. Typically, the chloroplast and pyrenoid are located on the bulging posterior end of the cell while the nucleus is located towards the apical end. Maintaining this shape during subsequent divisions should naturally force newly formed cells to arrange in a fan-shape, which if division continues will force the organism into a spherical form. Such a mechanism can rather simply define how

organismal polarity in the 4-celled *T. socialis* occurs – as cells divide they retain an apical to posterior polarity and hence cell shape defines organismal polarity. Interestingly, in *G. pectorale* it appears that at the 8-celled stage of embryogenesis, inversion actually forces the spherical form to open up into the flattened bowl shape characteristic of the mature *G. pectorale* form – suggesting an active process. In *G. pectorale* more "droplet-shaped" cells are formed on the inner top of the bowl while slightly rounder cells move to the outer bottom of the bowl. The asymmetry of cell shape appears therefore to signal for the outer peripheral cells to arrange flagella in a parallel fashion whilst the inner droplet-shaped cells form flagella in anti-parallel fashion and hence central-peripheral organismal polarity is attained. It is possible that in the simpler taxa that cell shape is a common mechanism for defining organism level polarity.

In the more complex *Volvox* species organismal polarity is defined as anterior to posterior (AP) polarity where gonidia are more common in the posterior and essentially absent from the most anterior regions of the organism. This raises the question of whether this AP polarity is derived from asymmetrical division? The AP polarity may arise from somatic cells moving further away from the site of division than what gonidia do and as such the AP polarity may not be "programmed" but rather a simple consequence of embryology. Possible factors that refute this theory include that additional traits also follow AP polarity, such as the observation that a gradient in somatic cell size and eyespot size exists

where larger somatic cells with larger eyespots are located more towards the anterior (with smaller somatic cells more towards the posterior) and that eyespots tend to be more towards the apical end of the cell in anterior cells than in posterior cells. Again, it is possibly that the cell shape of dividing cells could explain these phenomena but tracing cell divisions through the development of a ~2000-celled organism and following the cells during inversion would be extremely challenging.

10.5 Transformation of the Cell wall into an ECM and the Expansion of the ECM

Comparative genomics has revealed certain molecular underpinnings of this trait. The cell wall GP1 is absent from the multicellular taxa, including in *T. socialis*, which suggests that the loss of this gene may play a role in the transformation of the cell wall into an ECM. While the expansion of the ECM is underpinned by the copy number of ECM genes - *C. reinhardtii*, *T. socialis* and *G. pectorale* all possess similar numbers of ECM proteins while *V. carteri*, which is composed of >95 ECM, has a double the number of ECM genes. As opposed to a polygenic origin for *V. carteri* pherophorin and MMP gene expansion protein phylogenies show that the expansions were derived from only a smaller number of ancestral genes. A number of *V. carteri* genes display elevated omega values (dN/dS), which further points towards an adaptive role for these genes (Hanschen, et al. 2016). In the current analyses of

pherophorin and MMP proteins that include *T. socialis*, the same sets of curated pherophorin and MMP proteins from *C. reinhardtii*, *G. pectorale* and *V. carteri* were used as in Hanschen et al (2016). With the inclusion of *T. socialis* proteins both phylogenies are largely consistent with the phylogenies reported by Hanschen et al (2016). An interesting aspect of pherophorin and MMP gene duplication in *V. carteri* is that the pattern of duplication for both sets of proteins is similar.

This pattern is found in two ways: (i) the number of MMPs to pherophorins correlates throughout the lineage, and (ii) both pherophorin and MMP expansion in *V. carteri* is characterised by expansion from a limited number of ancestral genes. These two observations support the hypothesis of Heitzer and Hallman (2002) that specific MMPs might exist to target specific pherophorins. This speculation still requires further examination but to date genomic data seem to support this hypothesis.

10.6 Genetic Control of Cell number

Previous analyses and this study did not identify differences in the presence or absence of cyclin dependant kinases (CDKs) that may be associated with the evolution of multicellularity; the only identifiable change being the gain of *CDKG2* in *C. reinhardtii* or, alternatively, the loss of this gene in the multicellular taxa. Known modifications in the cell cycle include the expansion of D-type cyclins and the modification of the *rb* gene. The control of cell size and cell number in relation to the

multicellular forms is not as yet well understood. Although transferal of *G. pectorale* rb into *C. reinhardtii* can induce coloniality the mechanism for this transformation is as yet unknown. A maximum cell number of 16 for *rb*-transformed *C. reinhardtii* was identified by Hanschen et al (2016) but the deviation from this cell number was not reported and hence it is unknown if the cell number is controlled in these transformed cultures. Laboratory grown colonial volvocines stereotypically reach the maximum cell number characteristic of each species, which is not highly variable.

The absolute number of cells is in large part defined by the size of the mother cell because of the mode of cell division, which in the volvocines is by multiple fission division. For *Volvox* spp this may require the mother cell to reach a sufficient size for up to 16 rounds of division, whereas for *T. socialis* only two rounds of division are required to create 4-celled progeny. The genetic mechanism that determines the number of divisions characteristic of each species is unknown. In *C. reinhardtii* the *CDKG1* gene appears to be important for controlling the number of divisions (Li, et al. 2016), however, explanatory differences in the CDKG1 protein sequence could not be identified in the current investigation.

10.6.1 Cyclins

The evidence from Prochnik et al (2010) that *V. carteri* possesses 4-copies of cyclin D1 (in relation to a single copy in *C. reinhardtii*) was highlighted as a major finding of the genome analysis of *V. carteri*.

Furthermore, the *G. pectorale* and *T. socialis* genome sequence indicated that this expansion was present in this colonial organism and thereby correlated with the evolution of coloniality. The exact role that the expansion of cyclin D1 genes plays is as yet unknown but has been speculated to be associated with the retinoblastoma pathway (Hanschen, et al. 2016).

10.6.2 Retinoblastoma Gene

By including RB proteins from Hiraide et al (2013), more distantly related unicellular chlorophytes as well as the *T. socialis* RB protein the current study re-examined the role of the *rb* gene in the evolution of volvocine multicellularity. By including RB proteins from a larger subset of taxa the difference between multicellular and unicellular RB protein structure were less definitive than identified by Hanschen et al (2016). The RB-A to RB-B linker region was shown to be highly variable in all chlorophytes examined; regions longer and shorter than 100aa residues were found in both unicellular and multicellular chlorophytes. However, with the exception of *Yamagishiella unicocca* (106aa residues) the length of the linker region in the multicellular volvocines is shorter than 100aa residues and it is indeed longer in *C. reinhardtii* (130aa residues). If the RB protein has been co-opted from its ancestral unicellular function to play a critical role in the evolution of multicellularity then a reduction of 24aa residues would appear to be sufficient. Notably, the length of this linker region in other related unicellular organisms may also be shorter than

100aa such as in *Chlamydomonas chlamydogama*. Intriguingly, of all examined Chlorophyte RB protein sequences the *G. pectorale* RB protein has the shortest linker region between the RB-A and RB-B domains. This raises the possibility that the *G. pectorale* gene may unique in its transformative induction of *C. reinhardtii* colony formation. It would be interesting to transform *MAT3* deficient *C. reinhardtii* with the *Y. unicocca RB* gene to see if the same effect is observed.

In the current study the presence of phosphorylation targets within the linker region between RB-B and the C-terminal domain was not found to correlate with volvocine multicellularity. The typical CDK phosphorylation target motif is: (S/T)PX(K/R) with S/T each ~50% prevalent, P 100% prevalent, X indicates any amino acid and K/R are the most common but others may be present at less than 3% prevalence (Chang, et al. 2007). Typical CDK phosphorylation motifs within the volvocine RB-B/C-terminal linker region were identified in *C. reinhardtii*, *T. socialis*, *G. pectorale* and *V. carteri f nagariensis* (female – not previously identified) and in all proteins a single motif was identified. In addition, up to three CDK phosphorylation motifs of lower confidence (the 4^{th} base position is not a K or R) were identified for *C. reinhardtii*, *T. socialis*, *P. starrii*, *V. africanus*, *V. carteri f nagariensis* (female), *V. carteri f weismannia* (male) and *Eudorina* spp. Differences between the *G. pectorale* RB protein sequence generated by Hiraide et al (2013) and that identified by Hanschen et al (2016) were identified. An additional 12 AA

residues with the RB-B/C-terminal linker are present in the Hiraide et al (2013) RB protein sequence and a CDK phosphorylation motif (SPSK) is present within these residues. It is unknown which of these two sequences is most accurate.

Overall, the above indicates that the transformative ability of the *G. pectorale rb* gene is unlikely to be due to phosphorylation motifs in the RB-B/C-terminal linker region. The length of the RB-A and RB-B linker region is more likely to be of importance - possibly by altering protein-protein interactions or association with chromatin. However, phosphorylation targets have been identified purely by informatic prediction. Before *rb* is stated as a key factor in the development of multicellularity outstanding issues such as differential phosphorylation targets, the ability of RB proteins from multicellular taxa to bind E2F and DP, chromatin interactions and the effect of RB transformation from a series of volvocine algae must be addressed. Notwithstanding these issues, the effect of transforming *C. reinhardtii* into a group with a single gene transfer is remarkable.

10.6.3 Additional regulators of the cell cycle

As already mentioned key actors in the volvocine cell cycle are known but feeder pathways and co-regulatory mechanisms are less well characterised. Despite the clearly extensive differences in cell cycle control and the demonstrably heritable nature of cell cycle regulation in the volvocines the core cell cycle genes are highly conserved throughout the

volvocine lineage with only a few known exceptions, such as cyclin D1 expansion.

The limited differences of core cell cycle machinery possibly reflects at least 4 alternative scenarios for explaining the genetic modification of the cell cycle: (i) the co-option of ancestral core cell cycle genes (e.g. *MAT3* and *CYCD1*) into new roles is sufficient for explaining the genetic difference though all forms of volvocines, (ii) the regulation of core cell cycle genes through activation and degradation is most critical, (iii) the evolution of ancillary pathways that regulate core cell cycle genes explain cell cycle modifications or (iv) a combination of all the above. With current evidence it is impossible to discern between these alternatives. The absolute limit of the number of cell divisions in colonial volvocines suggests that a strict control unique to each species exists that controls either the entry into the cell cycle or the concentration of “sizer” molecules. A protein implicated in the “commitment” phase of cell cycle regulation is MAT/RB protein. However, the protein sequence of MAT3 alone seems unlikely as a candidate for controlling the number of cell divisions because identifiable variations in RB sequence structure, such as the length of the RB-A to RB-B linker region, do not correlate with the number of cell divisions characteristic of various volvocine species. A simple refutation of RB sequence structure as the dominant determinant of cell division number is the difference between *Volvox* male and female RB protein sequence structure if the length of the linker region is critical in how RB

acts then how do both and female *Volvox* strains divide a similar number of times during vegetative reproduction?

Although the experimental evidence points to a key role for RB in the lineage, the sequence composition seems insufficiently explanatory. This suggests that the regulation of RB and associated partners/pathways as being instrumental for the evolution of multicellularity rather than the innate sequence composition. Studies in metazoa, yeast and plants have identified a myriad of ancillary pathways and genes involved in regulating the cell cycle, however, similar ancillary participants in chlorophytes have not been the focus of many studies. In chapter 3 the role of the ubiquitin proteasomal pathway was highlighted as a pathway that qualifies as an 'additional regulator of the cell cycle'. Furthermore, DNA repair genes and numerous proteins kinases were identified as gained or under altered selection pressures in the multicellular lineage. Given the similarity of core cell cycle genes a thorough investigation of other potential regulators is warranted.

10.7 Germ-soma division of labour, Asymmetric Division and the Bifurcation of the Cell Division Program

The *regA* cluster of genes (Hanschen, et al. 2014; Grochau-Wright, et al. 2017) was not identified in *T. socialis* and is not present in *G.*

pectorale. Recent analyses point towards the origin of the *regA* cluster being near the origin of the Volvocaceae family, even though not all members of the Volvocaceae family possess germ-soma division of labour (Grochau-Wright, et al. 2017).

This indicates that other genomic innovations and/or a difference in the genotype/phenotype map produces germ-soma division of labour in some taxa but not in others. Genomic factors and interactions that may be important for the evolution of germ-soma division of labour cannot as yet be teased out without genomic data from Volvocaceae that do not possess germ-soma division of labour. Differences in gene expression likely play a large role in the division of germ and soma, hwoever, these differences are only known for *V. carteri*. (Matt and Umen 2018). From the current analysis the genome of *T. socialis* was not appropriate to provide insight into the evolution of germ-soma division and this was not the aim of the research. The same may be said for asymmetrical division and the bifurcation of the cell division program, which are traits present in more morphological complex taxa. It is worth noting however that extended cell cycle regulators may be important for understanding the bifurcation of the cell cycle. When more taxa are genome sequenced further understanding of these traits might be elucidated through genomic comparisons within the Volvocaceae.

APPENDIX E: TURNITIN REPORT

320265:Final_thesisTurnitin.docx

ORIGINALITY REPORT

10%	7%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jonathan Featherston, Yoko Arakaki, Erik R Hanschen, Patrick J Ferris et al. "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.", Molecular Biology and Evolution, 2017 Publication	3%
2	Featherston, Jonathan, Yoko Arakaki, Hisayoshi Nozaki, Pierre M. Durand, and David R. Smith. "Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in volvocine green algae", European Journal of Phycology, 2016. Publication	1%
3	www.mgh.de Internet Source	<1%
4	www.ncbi.nlm.nih.gov Internet Source	<1%
5	www.eebweb.arizona.edu	

	Internet Source	<1%
6	www.g3journal.org Internet Source	<1%
7	"Evolutionary Transitions to Multicellular Life", Springer Nature, 2015 Publication	<1%
8	www.webmedcentral.com Internet Source	<1%
9	www.plantcell.org Internet Source	<1%
10	pesquisa.bvsalud.org Internet Source	<1%
11	Yoko Arakaki, Takayuki Fujiwara, Hiroko Kawai-Toyooka, Kaoru Kawafune et al. "Evolution of cytokinesis-related protein localization during the emergence of multicellularity in volvocine green algae", BMC Evolutionary Biology, 2017 Publication	<1%
12	cshperspectives.cshlp.org Internet Source	<1%
13	edepot.wur.nl Internet Source	<1%
	wiredspace.wits.ac.za	

14	Internet Source	<1%
15	arizona.pure.elsevier.com Internet Source	<1%
16	biography.omicsonline.org Internet Source	<1%
17	biorxiv.org Internet Source	<1%
18	espace.library.uq.edu.au Internet Source	<1%
19	journals.plos.org Internet Source	<1%
20	peerj.com Internet Source	<1%
21	link.springer.com Internet Source	<1%
22	genome.cshlp.org Internet Source	<1%
23	eebweb.arizona.edu Internet Source	<1%
24	fmedicine.ajums.ac.ir Internet Source	<1%
25	Sa Geng, Ayano Miyagi, James G. Umen. "	<1%

	Evolutionary divergence of the sex-determining gene uncoupled from the transition to anisogamy in volvocine algae ", Development, 2018 Publication	
26	epub.uni-regensburg.de Internet Source	<1%
27	www.nature.com Internet Source	<1%
28	www.vliz.be Internet Source	<1%
29	"Chlamydomonas: Biotechnology and Biomedicine", Springer Nature, 2017 Publication	<1%
30	volvocales.pbworks.com Internet Source	<1%
31	www.plantphysiol.org Internet Source	<1%
32	Jonathan Featherston, Yoko Arakaki, Erik R Hanschen, Patrick J Ferris et al. "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", Molecular Biology and Evolution, 2017 Publication	<1%

33	www.imafungus.org Internet Source	<1%
34	ediss.uni-goettingen.de Internet Source	<1%
35	intl.plantcell.org Internet Source	<1%
36	eprints.qut.edu.au Internet Source	<1%
37	Sexual Reproduction in Animals and Plants, 2014. Publication	<1%
38	genomea.asm.org Internet Source	<1%
39	paperity.org Internet Source	<1%
40	philpapers.org Internet Source	<1%
41	edoc.ub.uni-muenchen.de Internet Source	<1%
42	repository.kulib.kyoto-u.ac.jp Internet Source	<1%
43	pdfs.semanticscholar.org Internet Source	<1%
	ec.asm.org	

44	Internet Source	<1%
45	pub.uni-bielefeld.de Internet Source	<1%
46	www.pomics.com Internet Source	<1%
47	carollee.labs.wisc.edu Internet Source	<1%
48	wallaby.vu.edu.au Internet Source	<1%
49	171.66.122.39 Internet Source	<1%
50	www.unbf.ca Internet Source	<1%
51	www.readbag.com Internet Source	<1%
52	Geng, Sa, Peter De Hoff, and James G. Umen. "Evolution of Sexes from an Ancestral Mating-Type Specification Pathway", PLoS Biology, 2014. Publication	<1%
53	Smith, D. R., T. Hamaji, B. J. S. C. Olson, P. M. Durand, P. Ferris, R. E. Michod, J. Featherston, H. Nozaki, and P. J. Keeling. "Organelle	<1%

	Genome Complexity Scales Positively with Organism Size in Volvocine Green Algae", Molecular Biology and Evolution, 2013. Publication	
54	www.slideshare.net Internet Source	<1%
55	arizona.openrepository.com Internet Source	<1%
56	dev.biologists.org Internet Source	<1%
57	ftp.dna.affrc.go.jp Internet Source	<1%
58	www.nap.edu Internet Source	<1%
59	www.mycologia.org Internet Source	<1%
60	Advances in Photosynthesis and Respiration, 1998. Publication	<1%
61	University of Tennessee, Knoxville Publication	<1%
62	mbe.library.arizona.edu Internet Source	<1%
63	jeb.biologists.org Internet Source	<1%

64	www.botany.wisc.edu Internet Source	<1%
65	theses.ulaval.ca Internet Source	<1%
66	imgpublic.mci-group.com Internet Source	<1%
67	www.plosgenetics.org Internet Source	<1%
68	mcb.berkeley.edu Internet Source	<1%
69	Herron, M. D.. "Origins of Multicellular Complexity: Volvox and the Volvocine Algae", Molecular Ecology, 2016. Publication	<1%
70	www.qucosa.de Internet Source	<1%
71	Robert Boyd, Joan B Silk. "How Humans Evolved 8E", W. W. Norton & Company, 2017 Publication	<1%
72	The Physiology of Microalgae, 2016. Publication	<1%
73	Matthew D. Herron. "EVOLUTION OF COMPLEXITY IN THE VOLVOCINE ALGAE: TRANSITIONS IN INDIVIDUALITY THROUGH	<1%

	DARWIN'S EYE", Evolution, 2/2008 Publication	
74	Efetova, Marina. "Molekulare Mechanismen einer wechselseitigen Kontrolle der Arabidopsis-Agrobacterium-Interaktion", Universität Würzburg, 2008. Publication	<1%
75	www.mdpi.com Internet Source	<1%
76	"New Peptides and Proteins Study Results from S. Belmabrouk et al Described (Exploring proteome-wide ", Life Science Weekly, July 12 2016 Issue Publication	<1%
77	Gavriel Y. Matt, James G. Umen. " Cell-Type Transcriptomes of the Multicellular Green Alga Yield Insights into the Evolutionary Origins of Germ and Somatic Differentiation Programs ", G3: Genes|Genomes|Genetics, 2018 Publication	<1%
78	digital.library.adelaide.edu.au Internet Source	<1%
79	intl.genetics.org Internet Source	<1%
80	as.vanderbilt.edu Internet Source	<1%

81	Hamaji, Takashi, David R. Smith, Hideki Noguchi, Atsushi Toyoda, Masahiro Suzuki, Hiroko Kawai-Toyooka, Asao Fujiyama, Ichiro Nishii, Tara Marriage, Bradley J. S. C. Olson, and Hisayoshi Nozaki. "Mitochondrial and Plastid Genomes of the Colonial Green Alga Gonium pectorale Give Insights into the Origins of Organelle DNA Architecture within the Volvocales", PLoS ONE, 2013. Publication	<1%
82	citeseerx.ist.psu.edu Internet Source	<1%
83	Armin Hallmann. "Evolution of reproductive development in the volvocine algae", Sexual Plant Reproduction, 12/21/2010 Publication	<1%
84	65.199.186.23 Internet Source	<1%
85	michodlab.arizona.edu Internet Source	<1%
86	kups.ub.uni-koeln.de Internet Source	<1%
87	www.welonsell.info Internet Source	<1%
88	gbe.oxfordjournals.org Internet Source	<1%

89	gigascience.biomedcentral.com Internet Source	<1%
90	aura.abdn.ac.uk Internet Source	<1%
91	mbe.oxfordjournals.org Internet Source	<1%
92	ecco.vub.ac.be Internet Source	<1%
93	run.unl.pt Internet Source	<1%
94	science.sciencemag.org Internet Source	<1%
95	www.springerlink.com Internet Source	<1%
96	www.pnas.org Internet Source	<1%
97	prirodni-akvarium.cz Internet Source	<1%
98	damtp.cam.ac.uk Internet Source	<1%
99	www.sajs.co.za Internet Source	<1%
100	elifesciences.org Internet Source	<1%

101	www.db-thueringen.de Internet Source	<1%
102	era.library.ualberta.ca Internet Source	<1%
103	www.phycologia.org Internet Source	<1%
104	darchive.mblwhoilibrary.org:8080 Internet Source	<1%
105	Clark Spencer Larsen. "Our Origins, Fourth Editon", W. W. Norton & Company, 2017 Publication	<1%
106	bio.biologists.org Internet Source	<1%
107	rnajournal.cshlp.org Internet Source	<1%
108	dalspace.library.dal.ca Internet Source	<1%
109	www.454.com Internet Source	<1%
110	qmro.qmul.ac.uk Internet Source	<1%
111	www.plosone.org Internet Source	<1%

112	epublications.uef.fi Internet Source	<1%
113	bmcbiol.biomedcentral.com Internet Source	<1%
114	rstb.royalsocietypublishing.org Internet Source	<1%
115	biology.mcgill.ca Internet Source	<1%
116	www.genetics.org Internet Source	<1%
117	etd.lsu.edu Internet Source	<1%
118	www.biologydirect.com Internet Source	<1%
119	g3journal.org Internet Source	<1%
120	www.php.wur.nl Internet Source	<1%
121	Lu, Yinghong(Hegemann, Peter, Lockau, Wolfgang and Mittag, Maria). "Functional analysis of phototropin in Chlamydomonas reinhardtii", Mathematisch-Naturwissenschaftliche Fakultät I, 2006. Publication	<1%

122	docnum.univ-lorraine.fr Internet Source	<1%
123	summit.sfu.ca Internet Source	<1%
124	www.fr.embnet.org Internet Source	<1%
125	irgc-portugal.org Internet Source	<1%
126	orbilu.uni.lu Internet Source	<1%
127	intl.plantphysiol.org Internet Source	<1%
128	repository.lib.ncsu.edu Internet Source	<1%
129	pure.uva.nl Internet Source	<1%
130	jcs.biologists.org Internet Source	<1%
131	bmcdevbiol.biomedcentral.com Internet Source	<1%
132	www.scielo.br Internet Source	<1%
133	www.gen.wur.nl Internet Source	

		<1%
134	www.dbbe.fcen.uba.ar Internet Source	<1%
135	www.beanblossom.in.us Internet Source	<1%
136	134.184.131.111 Internet Source	<1%
137	David L. Kirk. "A twelve-step program for evolving multicellularity and a division of labor", BioEssays, 03/2005 Publication	<1%
138	Matt, Gavriel, and James Umen. "Volvox: A simple algal model for embryogenesis, morphogenesis and cellular differentiation", Developmental Biology, 2016. Publication	<1%
139	Sonya Bahar. "The Essential Tension", Springer Nature, 2018 Publication	<1%
140	F. Ender. "Evidence for Autocatalytic Cross-Linking of Hydroxyproline-Rich Glycoproteins during Extracellular Matrix Assembly in Volvox", THE PLANT CELL ONLINE, 05/01/2002 Publication	<1%

APPENDIX F: ETHICS WAIVER

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
7 York Road, Parktown, 2193 South Africa * Telephone (011)717-2363

20th June2016

To: Whom it may concern,

Re: Waiver from Animal Ethics Screening Committee of the University of the Witwatersrand

This letter is to confirm that Dr. Pierre Durand (department of molecular medicine and haematology, university of the Witwatersrand) and his postgraduate students/Postdoctoral fellows do not require full animal ethics clearance to undertake the studies detailed below:

1. A comparative metagenomic analysis of a harmful algal bloom. Andrew Ndhlovu (PhD).
2. The Organelle Genomes of *Tetrabaena socialis*. Jonathan Featherston (PhD).
3. Evolution of programmed cell death in the unicellular chlorophyte *Chlamydomonas reinhardtii*. Nisha Dhar (postdoc).
4. Ecology and evolution of programmed death in bloom forming algae on the west coast of South Africa. Santsoh Sathe (postdoc).

No animals will be used in the projects which are either computational, mathematical, genomics or make use of microalgae algae from samples collected in South Africa.

If you would like any further or more specific information in this regard, do not hesitate to contact me.

Yours sincerely,



Kennedy Erlwanger
(Chairman: Animal Ethics Screening Committee, University of the Witwatersrand)

Reference: P Durand Waiver 20-06-2016-O

Assoc Prof Kennedy H. Erlwanger
School of Physiology
Faculty of Health Sciences, University of the Witwatersrand
7 York Road, Parktown, 2193
SOUTH AFRICA

Private bag 3, Wits, 2050, South Africa.
Tel: +27 (0)11 717 2454
Email: Kennedy.Erlwanger@wits.ac.za

REFERENCES

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Adair WS, Steinmetz SA, Mattson DM, Goodenough UW, Heuser JE. 1987. Nucleated assembly of Chlamydomonas and Volvox cell walls. J Cell Biol 105:2373-2382.

Adami C. 2002. What is complexity? Bioessays 24:1085-1094.

Adami C, Ofria C, Collier TC. 2000. Evolution of biological complexity. Proc Natl Acad Sci U S A 97:4463-4468.

Aktipis CA, Boddy AM, Jansen G, Hibner U, Hochberg ME, Maley CC, Wilkinson GS. 2015. Cancer across the tree of life: cooperation and cheating in multicellularity. Philos Trans R Soc Lond B Biol Sci 370.

Aktipis CA, Kwan VS, Johnson KA, Neuberg SL, Maley CC. 2011. Overlooking evolution: a systematic analysis of cancer relapse and therapeutic resistance research. PLoS One 6:e26100.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Amon P, Haas E, Sumper M. 1998. The sex-inducing pheromone and wounding trigger the same set of genes in the multicellular green alga Volvox. Plant Cell 10:781-789.

Arakaki Y, Kawai-Toyooka H, Hamamura Y, Higashiyama T, Noga A, Hirono M, Olson BJ, Nozaki H. 2013. The simplest integrated multicellular organism unveiled. PLoS One 8:e81641.

Araki N, Uchida A, Kimura T, Yoshikawa H, Aoki Y, Ueda T, Takai S, Miki T, Ono K. 1991. Involvement of the retinoblastoma gene in primary osteosarcomas and other bone and soft-tissue tumors. Clin Orthop Relat Res:271-277.

Armbrust EV, Ferris PJ, Goodenough UW. 1993. A mating type-linked gene cluster expressed in Chlamydomonas zygotes participates in the uniparental inheritance of the chloroplast genome. Cell 74:801-811.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman TL, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? PLoS Biol 12:e1001899.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-477.

Beardall J, Allen D, Bragg J, Finkel ZV, Flynn KJ, Quigg A, Rees TA, Richardson A, Raven JA. 2009. Allometry and stoichiometry of unicellular, colonial and multicellular phytoplankton. New Phytol 181:295-309.

Belsky JA, MacAlpine HK, Lubelsky Y, Hartemink AJ, MacAlpine DM. 2015. Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. Genes Dev 29:212-224.

Bernstein H, Byerly HC, Hopf FA, Michod RE. 1985a. The evolutionary role of recombinational repair and sex. Int Rev Cytol 96:1-28.

Bernstein H, Byerly HC, Hopf FA, Michod RE. 1985b. Genetic damage, mutation, and the evolution of sex. Science 229:1277-1281.
Bernstein H, Byerly HC, Hopf FA, Michod RE. 1984. Origin of sex. J Theor Biol 110:323-351.
Bisova K, Krylov DM, Umen JG. 2005. Genome-wide annotation and expression profiling of cell cycle regulatory genes in Chlamydomonas reinhardtii. Plant Physiol 137:475-491.
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 Plant Sci 19:672-680.
Blaby IK, Glaesener AG, Mettler T, Fitz-Gibbon ST, Gallaher SD, Liu B, Boyle NR, Kropat J, Stitt M, Johnson S, et al. 2013. Systems-level analysis of nitrogen starvation-induced modifications of carbon metabolism in a Chlamydomonas reinhardtii starchless mutant. Plant Cell 25:4305-4323.
Bodily PM, Fujimoto MS, Snell Q, Ventura D, Clement MJ. 2016. ScaffoldScaffolder: solving contig orientation via bidirected to directed graph reduction. Bioinformatics 32:17-24.
Boer PH, Bonen L, Lee RW, Gray MW. 1985. Genes for respiratory chain proteins and ribosomal RNAs are present on a 16-kilobase-pair DNA species from Chlamydomonas reinhardtii mitochondria. Proc Natl Acad Sci U S A 82:3340-3344.
Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578-579.
Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol 13:R56.
Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sorensen I, Lichtenstein G, et al. 2014. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet 46:1034-1038.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
Bonner JT. 1998. The Origins of Multicellularity. Intergrative Biology 1:27-36.
Boraas ME, D.B. S, Boxhorn JE. 1998. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. Evolutionary Ecology 12:153-164.
Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics Chapter 4:Unit 4 6 1-10.
Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. UniProtKB/Swiss-Prot. Methods Mol Biol 406:89-112.
Boyd M, Rosenzweig F, Herron MD. 2018. Analysis of motility in multicellular Chlamydomonas reinhardtii evolved under predation. PLoS One 13:e0192184.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2:10.
Brosius J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107:209-238.
Brumley DR, Polin M, Pedley TJ, Goldstein RE. 2015. Metachronal waves in the flagellar beating of Volvox and their hydrodynamic origin. J R Soc Interface 12:20141358.
Buttery NJ, Smith J, Queller DC, Strassmann JE. 2013. Measuring cheating, fitness, and segregation in Dictyostelium discoideum. Methods Mol Biol 983:231-248.
Campbell AM, Rayala HJ, Goodenough UW. 1995. The iso1 gene of Chlamydomonas is involved in sex determination. Mol Biol Cell 6:87-95.
Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.
Cavalier-Smith T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot 95:147-175.
Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci 34:247-278.
Cave MS, Pocock MA. 1951. Karyological Studies in the Volvocaceae. Am J Bot 38:800-811.
Chang EJ, Begum R, Chait BT, Gaasterland T. 2007. Prediction of cyclin-dependent kinase phosphorylation substrates. PLoS One 2:e656.
Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol 16:30.
Chen YM, Yu CH, Hwang CC, Liu T. 2013. OMACC: an Optical-Map-Assisted Contig Connector for improving de novo genome assembly. BMC Syst Biol 7 Suppl 6:S7.
Cheng Q, Fowler R, Tam LW, Edwards L, Miller SM. 2003. The role of GlsA in the evolution of asymmetric cell division in the green alga Volvox carteri. Dev Genes Evol 213:328-335.
Cheng Q, Pappas V, Hallmann A, Miller SM. 2005. Hsp70A and GlsA interact as partner chaperones to regulate asymmetric division in Volvox. Dev Biol 286:537-548.
Chu TC, Lu CH, Liu T, Lee GC, Li WH, Shih AC. 2013. Assembler for de novo assembly of large genomes. Proc Natl Acad Sci U S A 110:E3417-3424.
Chun KT, Mathias N, Goebl MG. 1996. Ubiquitin-dependent proteolysis and cell cycle control in yeast. Prog Cell Cycle Res 2:115-127.
Ciechanover A, Finley D, Varshavsky A. 1984. Ubiquitin dependence of selective protein degradation demonstrated in the mammalian cell cycle mutant ts85. Cell 37:57-66.

Colegrave N. 2012. The evolutionary success of sex. Science & Society Series on Sex and Science. EMBO Rep 13:774-778.
Coleman AW. 2012. A Comparative Analysis of the Volvocaceae (Chlorophyta)(1). J Phycol 48:491-513.
Coleman AW. 1999. Phylogenetic analysis of "Volvocacae" for comparative genetic studies. Proc Natl Acad Sci U S A 96:13892-13897.
Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 48:1193-1203.
Crespi B, Summers K. 2005. Evolutionary biology of cancer. Trends Ecol Evol 20:545-552.
Criscuolo A. 2011. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. Mol Phylogenet Evol 61:944-948.
Cross FR, Umen JG. 2015. The Chlamydomonas cell cycle. Plant J 82:370-392.
Cryns VL, Alexander JM, Klibanski A, Arnold A. 1993. The retinoblastoma gene in human pituitary tumors. J Clin Endocrinol Metab 77:644-646.
Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394-1403.
Dawkins R. 1976. The selfish gene. UK: Oxford University Press.
Dayarian A, Michael TP, Sengupta AM. 2010. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics 11:345.
Del Vasto M, Figueroa-Martinez F, Featherston J, Gonzalez MA, Reyes-Prieto A, Durand PM, Smith DR. 2015. Massive and widespread organelle genomic expansion in the green algal genus Dunaliella. Genome Biol Evol 7:656-663.
Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-375.
Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181-1184.
Didur O, Dewez D, Popovic R. 2013. Alteration of chromium effect on photosystem II activity in Chlamydomonas reinhardtii cultures under different synchronized state of the cell cycle. Environ Sci Pollut Res Int 20:1870-1875.
Diggle SP, Griffin AS, Campbell GS, West SA. 2007. Cooperation and conflict in quorum-sensing bacterial populations. Nature 450:411-414.
Dillard SR, Van K, Spalding MH. 2011. Acclimation to low or limiting CO2 in non-synchronous Chlamydomonas causes a transient synchronization of the cell division cycle. Photosynth Res 109:161-168.
Dinc E, Tian L, Roy LM, Roth R, Goodenough U, Croce R. 2016. LHCSR1 induces a fast and reversible pH-dependent fluorescence quenching in LHCII in Chlamydomonas reinhardtii cells. Proc Natl Acad Sci U S A 113:7673-7678.

Dohmen E, Kremer LP, Bornberg-Bauer E, Kemena C. 2016. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics 32:2577-2581.
Donnan L, John PC. 1983. Cell cycle control by timer and sizer in Chlamydomonas. Nature 304:630-633.
Drosophila 12 Genomes C, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-218.
Du X, Servin B, Womack JE, Cao J, Yu M, Dong Y, Wang W, Zhao S. 2014. An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae. BMC Genomics 15:625.
Dueck A, Evers M, Henz SR, Unger K, Eichner N, Merkl R, Berezikov E, Engelmann JC, Weigel D, Wenzl S, et al. 2016. Gene silencing pathways found in the green alga Volvox carteri reveal insights into evolution and origins of small RNA systems in plants. BMC Genomics 17:853.
Duncan L, Nishii I, Harryman A, Buckley S, Howard A, Friedman NR, Miller SM. 2007. The VARL gene family and the evolutionary origins of the master cell-type regulatory gene, regA, in Volvox carteri. J Mol Evol 65:1-11.
Duncan L, Nishii I, Howard A, Kirk D, Miller SM. 2006. Orthologs and paralogs of regA, a master cell-type regulatory gene in Volvox carteri. Curr Genet 50:61-72.
Ebel C, Mariconti L, Gruissem W. 2004. Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. Nature 429:776-780.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.
El Albani A, Bengtson S, Canfield DE, Bekker A, Macchiarelli R, Mazurier A, Hammarlund EU, Boulvais P, Dupuy JJ, Fontaine C, et al. 2010. Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago. Nature 466:100-104.
Ender F, Godl K, Wenzl S, Sumper M. 2002. Evidence for autocatalytic cross-linking of hydroxyproline-rich glycoproteins during extracellular matrix assembly in Volvox. Plant Cell 14:1147-1160.
Ender F, Hallmann A, Amon P, Sumper M. 1999. Response to the sexual pheromone and wounding in the green alga volvox: induction of an extracellular glycoprotein consisting almost exclusively of hydroxyproline. J Biol Chem 274:35023-35028.
Ertl H, Hallmann A, Wenzl S, Sumper M. 1992. A novel extensin that may organize extracellular matrix biogenesis in Volvox carteri. EMBO J 11:2055-2062.
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.

Featherston J, Arakaki Y, Nozaki H, Durand PM, Smith DR. 2016. Inflated organelle genomes and circular-mapping mtDNA probably existed at the origin of coloniality in volvocine green algae. European Journal of Phycology:1-9.
Featherston J, Durand PM. 2012. Cooperation and conflict in cancer: An evolutionary perspective. SAJS 108:1-12.
Fernandez A, Lynch M. 2011. Non-adaptive origins of interactome complexity. Nature 474:502-505.
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, 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, Goodenough UW. 1994. The mating-type locus of Chlamydomonas reinhardtii contains highly rearranged DNA sequences. Cell 76:1135-1145.
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.
Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 update. Nucleic Acids Res 43:W30-38.
Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ, Jr. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6:Unit 6 12 11-19.
Fisher RM, Bell T, West SA. 2016. Multicellular group formation in response to predators in the alga Chlorella vulgaris. J Evol Biol 29:551-559.
Fortunato A, Boddy A, Mallo D, Aktipis A, Maley CC, Pepper JW. 2017. Natural Selection in Cancer Biology: From Molecular Snowflakes to Trait Hallmarks. Cold Spring Harb Perspect Med 7.
Fortunato A, Strassmann JE, Santorelli L, Queller DC. 2003. Co-occurrence in nature of different clones of the social amoeba, Dictyostelium discoideum. Mol Ecol 12:1031-1038.
Gao S, Sung WK, Nagarajan N. 2011. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol 18:1681-1691.
Gardner A, West SA, Wild G. 2011. The genetical theory of kin selection. J Evol Biol 24:1020-1043.
Geng S, De Hoff P, Umen JG. 2014. Evolution of sexes from an ancestral mating-type specification pathway. PLoS Biol 12:e1001904.

Gioti A, Mushegian AA, Strandberg R, Stajich JE, Johannesson H. 2012. Unidirectional evolutionary transitions in fungal mating systems and the role of transposable elements. Mol Biol Evol 29:3215-3226.
Gisselsson D, Hakanson U, Stoller P, Marti D, Jin Y, Rosengren AH, Stewenius Y, Kahl F, Panagopoulos I. 2008. When the genome plays dice: circumvention of the spindle assembly checkpoint and near-random chromosome segregation in multipolar cancer cell mitoses. PLoS One 3:e1871.
Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108:1513-1518.
Godl K, Hallmann A, Rappel A, Sumper M. 1995. Pherophorins: a family of extracellular matrix glycoproteins from Volvox structurally related to the sex-inducing pheromone. Planta 196:781-787.
Godl K, Hallmann A, Wenzl S, Sumper M. 1997. Differential targeting of closely related ECM glycoproteins: the pherophorin family from Volvox. EMBO J 16:25-34.
Goodenough U. 2015. Historical perspective on Chlamydomonas as a model for basic research: 1950-1970. Plant J 82:365-369.
Goodenough UW, Forest CL. 1978. Genetics of cell-cell interactions in Chlamydomonas reinhardi. Birth Defects Orig Artic Ser 14:429-438.
Goodenough UW, Gebhart B, Mecham RP, Heuser JE. 1986. Crystals of the Chlamydomonas reinhardtii cell wall: polymerization, depolymerization, and purification of glycoprotein monomers. J Cell Biol 103:405-417.
Goodenough UW, Heuser JE. 1985. The Chlamydomonas cell wall and its constituent glycoproteins analyzed by the quick-freeze, deep-etch technique. J Cell Biol 101:1550-1568.
Goodenough UW, Levine RP. 1974. Methods for the study of chloroplast membranes of wild-type and mutant strains of Chlamydomonas reinhardi. Methods Enzymol 32:871-880.
Gould SJ. 1996. Full House: The Spread of Excellence from Plato to Darwin. USA: Harmony Books.
Green KJ, Kirk DL. 1981. Cleavage patterns, cell lineages, and development of a cytoplasmic bridge system in Volvox embryos. J Cell Biol 91:743-755.
Green KJ, Kirk DL. 1982. A revision of the cell lineages recently reported for Volvox carteri embryos. J Cell Biol 94:741-742.
Green KJ, Viamontes GI, Kirk DL. 1981. Mechanism of formation, ultrastructure, and function of the cytoplasmic bridge system during morphogenesis in Volvox. J Cell Biol 91:756-769.
Grochau-Wright ZI, Hanschen ER, Ferris PJ, Hamaji T, Nozaki H, Olson B, Michod RE. 2017. Genetic basis for soma is present in undifferentiated volvocine green algae. J Evol Biol 30:1205-1218.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript

sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494-1512.
Haas PA, Goldstein RE. 2015. Elasticity and glocality: initiation of embryonic inversion in Volvox. J R Soc Interface 12.
Halanych KM, Whelan NV, Kocot KM, Kohn AB, Moroz LL. 2016. Miscues misplace sponges. Proc Natl Acad Sci U S A 113:E946-947.
Hallmann A. 1999. Enzymes in the extracellular matrix of Volvox: an inducible, calcium-dependent phosphatase with a modular composition. J Biol Chem 274:1691-1697.
Hallmann A. 2011. Evolution of reproductive development in the volvocine algae. Sex Plant Reprod 24:97-112.
Hallmann A. 2003. Extracellular matrix and sex-inducing pheromone in Volvox. Int Rev Cytol 227:131-182.
Hallmann A, Amon P, Godl K, Heitzer M, Sumper M. 2001. Transcriptional activation by the sexual pheromone and wounding: a new gene family from Volvox encoding modular proteins with (hydroxy)proline-rich and metalloproteinase homology domains. Plant J 26:583-593.
Hallmann A, Godl K, Wenzl S, Sumper M. 1998. The highly efficient sex-inducing pheromone system of Volvox. Trends Microbiol 6:185-189.
Hallmann A, Kirk DL. 2000. The developmentally regulated ECM glycoprotein ISG plays an essential role in organizing the ECM and orienting the cells of Volvox. J Cell Sci 113 Pt 24:4605-4617.
Hamaji T, Ferris PJ, Coleman AW, Waffenschmidt S, Takahashi F, Nishii I, Nozaki H. 2008. Identification of the minus-dominance gene ortholog in the mating-type locus of Gonium pectorale. Genetics 178:283-294.
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, Ferris PJ, Nishii I, Nozaki H. 2009. Identification of the Minus Mating-Type Specific Gene Mtd1 from Gonium Pectorale (Volvocales, Chlorophyta)(1). J Phycol 45:1310-1314.
Hamaji T, Kawai-Toyooka H, Toyoda A, Minakuchi Y, Suzaki M, Fujiiyama A, Nozaki H, Smith DR. 2017. Multiple Independent Change in Mitochondrial Genome Conformation in Chlamydomonadalean Algae. Genome Biol Evol 9:993-999.
Hamaji T, Mogi Y, Ferris PJ, Mori T, Miyagishima S, Kabeya Y, Nishimura Y, Toyoda A, Noguchi H, Fujiyama A, et al. 2016. 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.
Hamaji T, Smith DR, Noguchi H, Toyoda A, Suzuki M, Kawai-Toyooka H, Fujiyama A, Nishii I, Marriage T, Olson BJ, et al. 2013. Mitochondrial and plastid genomes of the colonial green alga Gonium pectorale give insights into the origins of organelle DNA architecture within the volvocales. PLoS One 8:e57177.
Hamilton WD. 1963. The evolution of altruistic behavior American Naturalist 97:354-356.

Hanschen ER, Davison DR, Grochau-Wright ZI, Michod RE. 2017. Evolution of individuality: A Case Study in the Volvocine Green Algae Philos Theor Ptract Biol 9.
Hanschen ER, Ferris PJ, Michod RE. 2014. Early evolution of the genetic basis for soma in the volvocaceae. Evolution 68:2014-2025.
Hanschen ER, Herron MD, Wiens JJ, Nozaki H, Michod RE. 2017. Repeated evolution and reversibility of self-fertilization in the volvocine green algae. Evolution.
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.
Harper JF, Huson KS, Kirk DL. 1987. Use of repetitive sequences to identify DNA polymorphisms linked to regA, a developmentally important locus in Volvox. Genes Dev 1:573-584.
Harris EH. 1989. The Chlamydomonas sourcebook : a comprehensive guide to biology and laboratory use. San Diego: Academic Press.
Hartfield M, Keightley PD. 2012. Current hypotheses for the evolution of sex and recombination. Integr Zool 7:192-209.
Hatje K, Hammesfahr B, Kollmar M. 2013. WebScipio: Reconstructing alternative splice variants of eukaryotic proteins. Nucleic Acids Res 41:W504-509.
Heitzer M, Hallmann A. 2002. An extracellular matrix-localized metalloproteinase with an exceptional QEXXH metal binding site prefers copper for catalytic activity. J Biol Chem 277:28280-28286.
Heng HH. 2007. Elimination of altered karyotypes by sexual reproduction preserves species identity. Genome 50:517-524.
Heng HH. 2009. The genome-centric concept: resynthesis of evolutionary theory. Bioessays 31:512-525.
Heng HH, Bremer SW, Stevens J, Ye KJ, Miller F, Liu G, Ye CJ. 2006. Cancer progression by non-clonal chromosome aberrations. J Cell Biochem 98:1424-1435.
Heng HH, Liu G, Bremer S, Ye KJ, Stevens J, Ye CJ. 2006. Clonal and non-clonal chromosome aberrations and genome variation and aberration. Genome 49:195-204.
Heng HH, Liu G, Stevens JB, Bremer SW, Ye KJ, Abdallah BY, Horne SD, Ye CJ. 2011. Decoding the genome beyond sequencing: the new phase of genomic research. Genomics 98:242-252.
Heng HH, Stevens JB, Bremer SW, Liu G, Abdallah BY, Ye CJ. 2011. Evolutionary mechanisms and diversity in cancer. Adv Cancer Res 112:217-253.
Heng HH, Stevens JB, Bremer SW, Ye KJ, Liu G, Ye CJ. 2010. The evolutionary mechanism of cancer. J Cell Biochem 109:1072-1084.
Heng HH, Stevens JB, Lawrenson L, Liu G, Ye KJ, Bremer SW, Ye CJ. 2008. Patterns of genome dynamics and cancer evolution. Cell Oncol 30:513-514.

Heng HH, Stevens JB, Liu G, Bremer SW, Ye KJ, Reddy PV, Wu GS, Wang YA, Tainsky MA, Ye CJ. 2006. Stochastic cancer progression driven by non-clonal chromosome aberrations. J Cell Physiol 208:461-472.
Herron MD. 2009. Many from one: Lessons from the volvocine algae on the evolution of multicellularity. Commun Integr Biol 2:368-370.
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.
Herron MD, Michod RE. 2008. Evolution of complexity in the volvocine algae: transitions in individuality through Darwin's eye. Evolution 62:436-451.
Herron MD, Rashidi A, Shelton DE, Driscoll WW. 2013. Cellular differentiation and individuality in the 'minor' multicellular taxa. Biol Rev Camb Philos Soc 88:844-861.
Hiraide R, Kawai-Toyooka H, Hamaji T, Matsuzaki R, Kawafune K, Abe J, Sekimoto H, Umen J, Nozaki H. 2013. The evolution of male-female sexual dimorphism predates the gender-based divergence of the mating locus gene MAT3/RB. Mol Biol Evol 30:1038-1040.
Hoelzer MA, Michod RE. 1991. DNA repair and the evolution of transformation in Bacillus subtilis. III. Sex with damaged DNA. Genetics 128:215-223.
Hogg A, Bia B, Onadim Z, Cowell JK. 1993. Molecular mechanisms of oncogenic mutations in tumors from patients with bilateral and unilateral retinoblastoma. Proc Natl Acad Sci U S A 90:7351-7355.
Hohn S, Hallmann A. 2011. There is more than one way to turn a spherical cellular monolayer inside out: type B embryo inversion in Volvox globator. BMC Biol 9:89.
Hohn S, Honerkamp-Smith AR, Haas PA, Trong PK, Goldstein RE. 2015. Dynamics of a Volvox embryo turning itself inside out. Phys Rev Lett 114:178101.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.
Howell SH, Posakony JW, Hill KR. 1977. The cell cycle program of polypeptide labeling in Chlamydomonas reinhardtii. J Cell Biol 72:223-241.
Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28:593-594.
Huber O, Sumper M. 1994. Algal-CAMs: isoforms of a cell adhesion molecule in embryos of the alga Volvox with homology to Drosophila fasciclin I. EMBO J 13:4212-4222.
Hubisz MJ, Pollard KS. 2014. Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. Curr Opin Genet Dev 29:15-21.
Hunt M, Newbold C, Berriman M, Otto TD. 2014. A comprehensive evaluation of assembly scaffolding tools. Genome Biol 15:R42.
Huskey RJ, Griffin BE. 1979. Genetic control of somatic cell differentiation in Volvox analysis of somatic regenerator mutants. Dev Biol 72:226-235.

Iida H, Ota S, Inouye I. 2013. Cleavage, incomplete inversion, and cytoplasmic bridges in Gonium pectorale (Volvocales, Chlorophyta). J Plant Res 126:699-707.
Inglis RF, Ryu E, Asikhia O, Strassmann JE, Queller DC. 2017. Does high relatedness promote cheater-free multicellularity in synthetic lifecycles? J Evol Biol 30:985-993.
Jaenicke L, Feldwisch O, Merkl B, Cremer A, Haas IG. 1993. Expression of highly active sex-inducing pheromone of Volvox carteri f. nagariensis in a mammalian cell system. FEBS Lett 316:257-260.
Janero DR, Barrnett R. 1981. Thylakoid membrane biogenesis in Chlamydomonas reinhardtii 137+: cell cycle variations in the synthesis and assembly of polar glycerolipid. J Cell Biol 91:126-134.
John PC. 1984. Control of the cell division cycle in Chlamydomonas. Microbiol Sci 1:96-101.
Kapsetaki SE, Fisher RM, West SA. 2016. Predation and the formation of multicellular groups in algae. Evol Ecol Res 17:651-669.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
Kawai-Toyooka H, Mori T, Hamaji T, Suzuki M, Olson BJ, Uemura T, Ueda T, Nakano A, Toyoda A, Fujiyama A, et al. 2014. Sex-specific posttranslational regulation of the gamete fusogen GCS1 in the isogamous volvocine alga Gonium pectorale. Eukaryot Cell 13:648-656.
Kianianmomeni A, Nematollahi G, Hallmann A. 2008. A gender-specific retinoblastoma-related protein in Volvox carteri implies a role for the retinoblastoma protein family in sexual development. Plant Cell 20:2399-2419.
King N, Westbrook J, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, et al. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451:783-788.
Kirk DL. 1999. Evolution of multicellularity in the volvocine algae. Curr Opin Plant Biol 2:496-501.
Kirk DL. 1997. The genetic program for germ-soma differentiation in Volvox. Annu Rev Genet 31:359-380.
Kirk DL. 1994. Germ cell specification in Volvox carteri. Ciba Found Symp 182:2-15; discussion 15-30.
Kirk DL. 2001. Germ-soma differentiation in volvox. Dev Biol 238:213-223.
Kirk DL. 1988. The ontogeny and phylogeny of cellular differentiation in Volvox. Trends Genet 4:32-36.
Kirk DL. 2006. Oogamy: inventing the sexes. Curr Biol 16:R1028-1030.
Kirk DL. 2003. Seeking the ultimate and proximate causes of volvox multicellularity and cellular differentiation. Integr Comp Biol 43:247-253.
Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor. Bioessays 27:299-310.
Kirk DL. 1998. Volvox : molecular-genetic origins of multicellularity and cellular differentiation. Cambridge, U.K. ; New York: Cambridge University Press.

Kirk DL, Baran GJ, Harper JF, Huskey RJ, Huson KS, Zagris N. 1987. Stage-specific hypermutability of the regA locus of Volvox, a gene regulating the germ-soma dichotomy. Cell 48:11-24.
Kirk DL, Birchem R, King N. 1986. The extracellular matrix of Volvox: a comparative study and proposed system of nomenclature. J Cell Sci 80:207-231.
Kirk DL, Harper JF. 1986. Genetic, biochemical, and molecular approaches to Volvox development and evolution. Int Rev Cytol 99:217-293.
Kirk DL, Kaufman MR, Keeling RM, Stamer KA. 1991. Genetic and cytological control of the asymmetric divisions that pattern the Volvox embryo. Dev Suppl 1:67-82.
Kirk DL, Nishii I. 2001. Volvox carteri as a model for studying the genetic and cytological control of morphogenesis. Dev Growth Differ 43:621-631.
Kirk MM, Kirk DL. 2004. Exploring germ-soma differentiation in Volvox. J Biosci 29:143-152.
Kirk MM, Ransick A, McRae SE, Kirk DL. 1993. The relationship between cell size and cell fate in Volvox carteri. J Cell Biol 123:191-208.
Kirk MM, Stark K, Miller SM, Muller W, Taillon BE, Gruber H, Schmitt R, Kirk DL. 1999. regA, a Volvox gene that plays a central role in germ-soma differentiation, encodes a novel regulatory protein. Development 126:639-647.
Kleiber M. 1932. Body size and metabolism. Hilgardia 6:315-353.
Knoll AH. 2011. The multiple origins of complex multicellularity. Annual Review of Earth and Planetary Sciences 39:217-239.
Kochert G. 1968. Differentiation of reproductive cells in Volvox carteri. J Protozool 15:438-452.
Kohn AB, Citarella MR, Kocot KM, Bobkova YV, Halanych KM, Moroz LL. 2012. Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, Pleurobrachia bachei. Mol Phylogenet Evol 63:203-207.
Koonin EV. 2011. The Logic of Chance: The Nature and Origin of Biological Evolution. Upper Saddle River: FT Press.
Koonin EV. 2016. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC Biol 14:114.
Koonin EV, Csuros M, Rogozin IB. 2013. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip Rev RNA 4:93-105.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, et al. 2010. The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420.
Kuck P, Meusemann K. 2010. FASconCAT: Convenient handling of data matrices. Mol Phylogenet Evol 56:1115-1118.
Kuzdzal-Fick JJ, Fox SA, Strassmann JE, Queller DC. 2011. High relatedness is necessary and sufficient to maintain multicellularity in Dictyostelium. Science 334:1548-1551.

Larson A, Kirk MM, Kirk DL. 1992. Molecular phylogeny of the volvocine flagellates. Mol Biol Evol 9:85-105.
Lee SC, Ni M, Li W, Shertz C, Heitman J. 2010. The evolution of sex: a perspective from the fungal kingdom. Microbiol Mol Biol Rev 74:298-340.
Leslie MP, Shelton DE, Michod RE. 2017. Generation time and fitness tradeoffs during the evolution of multicellularity. J Theor Biol 430:92-102.
Levin SR, Brock DA, Queller DC, Strassmann JE. 2015. Concurrent coevolution of intra-organismal cheaters and resisters. J Evol Biol 28:756-765.
Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al. 2002. Apollo: a sequence annotation editor. Genome Biol 3:RESEARCH0082.
Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-2189.
Li Y, Liu D, Lopez-Paz C, Olson BJSC, Umen JG. 2016. A new class of cyclin dependent kinase in Chlamydomonas is required for coupling cell size to cell division. Elife 10767.
Lin H, Goodenough UW. 2007. Gametogenesis in the Chlamydomonas reinhardtii minus mating type is controlled by two genes, MID and MTD1. Genetics 176:913-925.
Linardopoulos S, Gonos ES, Spandidos DA. 1993. Abnormalities of retinoblastoma gene structure in human lung tumors. Cancer Lett 71:67-74.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.
Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet 17:704-714.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.
Mages HW, Tschochner H, Sumper M. 1988. The sexual inducer of Volvox carteri. Primary structure deduced from cDNA sequence. FEBS Lett 234:407-410.
Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29:1718-1725.
Mahjoub MR, Montpetit B, Zhao L, Finst RJ, Goh B, Kim AC, Quarmby LM. 2002. The FA2 gene of Chlamydomonas encodes a NIMA family kinase with roles in cell cycle progression and microtubule severing during deflagellation. J Cell Sci 115:1759-1768.
Maley CC, Forrest S. 2000. Exploring the relationship between neutral and selective mutations in cancer. Artif Life 6:325-345.
Maliet O, Shelton DE, Michod RE. 2015. A model for the origin of group reproduction during the evolutionary transition to multicellularity. Biol Lett 11:20150157.
Marguerat S, Bahler J. 2012. Coordinating genome expression with cell size. Trends in Genetics 28:560-565.

Mariano DC, Sousa Tde J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TD, Dorella FA, et al. 2016. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. BMC Genomics 17:315.
Marriage T, Olson BJ editors. The Third International Volvox Conference. 2015 Cambridge, UK.
Marshall WB, Young KD, Swaffer M, Wood E, Nurse P, Kimura A, Frankel J, Wallingford J, Wlabot V, Qu X, et al. 2012. What determines cell size? BMC Biol 10:1-22.
Martin NC, Chiang KS, Goodenough UW. 1976. Turnover of chloroplast and cytoplasmic ribosomes during gametogenesis in Chlamydomonas reinhardi. Dev Biol 51:190-201.
Martin NC, Goodenough UW. 1975. Gametic differentiation in Chlamydomonas reinhardtii. I. Production of gametes and their fine structure. J Cell Biol 67:587-605.
Matsumura K, Yagi T, Yasuda K. 2003. Role of timer and sizer in regulation of Chlamydomonas cell cycle. Biochem Biophys Res Commun 306:1042-1049.
Matt G, Umen J. 2016. Volvox: A simple algal model for embryogenesis, morphogenesis and cellular differentiation. Dev Biol 419:99-113.
Matt GY, Umen JG. 2018. Cell-Type Transcriptomes of the Multicellular Green Alga Volvox carteri Yield Insights into the Evolutionary Origins of Germ and Somatic Differentiation Programs. G3 (Bethesda) 8:531-550.
Maul JE, Lilly JW, Cui L, Claude W, 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.
Maynard Smith J. 1978. The Evolution of Sex. Cambridge UK: Cambridge University Press.
Maynard Smith J, Szathmáry E. 1995. The major transitions in evolution. Oxford ; New York: W.H. Freeman Spektrum.
McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. 2014. Insights into three whole-genome duplications gleaned from the Paramecium caudatum genome sequence. Genetics 197:1417-1428.
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.
Merlo LM, Pepper JW, Reid BJ, Maley CC. 2006. Cancer as an evolutionary and ecological process. Nat Rev Cancer 6:924-935.
Merlo LM, Wang LS, Pepper JW, Rabinovitch PS, Maley CC. 2010. Polyploidy, aneuploidy and the evolution of cancer. Adv Exp Med Biol 676:1-13.
Metin B, Findley K, Heitman J. 2010. The mating type locus (MAT) and sexual reproduction of Cryptococcus heveanensis: insights into the evolution of sex and sex-determining chromosomal regions in fungi. PLoS Genet 6:e1000961.
Michaelis G, Vahrenholz C, Pratje E. 1990. Mitochondrial DNA of Chlamydomonas reinhardtii: the gene for apocytochrome b and the

complete functional map of the 15.8 kb DNA. Mol Gen Genet 223:211-216.
Michaels A, Herrin DL. 1990. Translational regulation of chloroplast gene expression during the light-dark cell cycle of Chlamydomonas: evidence for control by ATP/energy supply. Biochem Biophys Res Commun 170:1082-1088.
Michod RE. 1996. Cooperation and conflict in the evolution of individuality. II. Conflict mediation. Proc Biol Sci 263:813-822.
Michod RE. 2006. The group covariance effect and fitness trade-offs during evolutionary transitions in individuality. Proc Natl Acad Sci U S A 103:9113-9117.
Michod RE, Herron MD. 2006. Cooperation and conflict during evolutionary transitions in individuality. J Evol Biol 19:1406-1409; discussion 1426-1436.
Michod RE, Nedelcu AM. 2003. On the reorganization of fitness during evolutionary transitions in individuality. Integr Comp Biol 43:64-73.
Michod RE, Nedelcu AM, Roze D. 2003. Cooperation and conflict in the evolution of individuality. IV. Conflict mediation and evolvability in Volvox carteri. Biosystems 69:95-114.
Michod RE, Roze D. 2001. Cooperation and conflict in the evolution of multicellularity. Heredity (Edinb) 86:1-7.
Michod RE, Roze D. 1997. Transitions in individuality. Proc Biol Sci 264:853-857.
Michod RE, Viossat Y, Solari CA, Hurand M, Nedelcu AM. 2006. Life-history evolution and the origin of multicellularity. J Theor Biol 239:257-272.
Miller SM. 2002. Taming the fierce roller: an "enhanced" understanding of cellular differentiation in Volvox. Bioessays 24:3-7.
Miller SM, Kirk DL. 1999. glsA, a Volvox gene required for asymmetric division and germ cell specification, encodes a chaperone-like protein. Development 126:649-658.
Miller SM, Schmitt R, Kirk DL. 1993. Jordan, an active Volvox transposable element similar to higher plant transposons. Plant Cell 5:1125-1138.
Mogi Y, Hamaji T, Suzuki M, Ferris P, Mori T, Kabeya Y, Miyagishima SY, Nozaki H. 2012. Evidence for Tubular Mating Structures Induced in Each Mating Type of Heterothallic Gonium Pectorale (Volvocales, Chlorophyta)(1). J Phycol 48:670-674.
Moore LJ, Coleman AW. 1989. The linear 20 kb mitochondrial genome of Pandorina morum (Volvocaceae, Chlorophyta). Plant Mol Biol 13:459-465.
Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, et al. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature 510:109-114.
Muhlhausen S, Hellkamp M, Kollmar M. 2015. GenePainter v. 2.0 resolves the taxonomic distribution of intron positions. Bioinformatics 31:1302-1304.

Muller H, Lukas J, Schneider A, Warthoe P, Bartek J, Eilers M, Strauss M. 1994. Cyclin D1 expression is regulated by the retinoblastoma protein. Proc Natl Acad Sci U S A 91:2945-2949.
Murray AW. 2013. Multicellularity: Is There A Big Idea? In. KITP Multicell13. California: Kavli Institute for Theoretical Physics.
Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nat Rev Genet 14:157-167.
Nakada T, Misawa K, Nozaki H. 2008. Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Mol Phylogenet Evol 48:281-291.
Nakamura T, Iwamura Y, Kaneko M, Nakagawa K, Kawai K, Mitamura K, Futagawa T, Hayashi H. 1991. Deletions and rearrangements of the retinoblastoma gene in hepatocellular carcinoma, insulinoma and some neurogenic tumors as found in a study of 121 tumors. Jpn J Clin Oncol 21:325-329.
Narasimha AM, Kaulich M, Shapiro GS, Choi YJ, Sicinski P, Dowdy SF. 2014. Cyclin D activates the Rb tumor suppressor by mono-phosphorylation. Elife 3.
Nedelcu AM, Michod RE. 2006. The evolutionary origin of an altruistic gene. Mol Biol Evol 23:1460-1464.
Nieuwenhuis BP, Billiard S, Vuilleumier S, Petit E, Hood ME, Giraud T. 2013. Evolution of uni- and bifactorial sexual compatibility systems in fungi. Heredity (Edinb) 111:445-455.
Niklas KJ. 2014. The evolutionary-developmental origins of multicellularity. Am J Bot 101:6-25.
Niklas KJ. 2006. Plant allometry, leaf nitrogen and phosphorus stoichiometry, and interspecific trends in annual growth rates. Ann Bot 97:155-163.
Niklas KJ. 2004. Plant allometry: is there a grand unifying theory? Biol Rev Camb Philos Soc 79:871-889.
Niklas KJ, Cobb ED, Crawford DR. 2013. The evo-devo of multinucleate cells, tissues, and organisms, and an alternative route to multicellularity. Evol Dev 15:466-474.
Niklas KJ, Kutschera U. 2015. Kleiber's Law: How the Fire of Life ignited debate, fueled theory, and neglected plants as model organisms. Plant Signal Behav 10:e1036216.
Niklas KJ, Newman SA. 2013. The origins of multicellular organisms. Evol Dev 15:41-52.
Niklas KJ, Spatz HC. 2004. Growth and hydraulic (not mechanical) constraints govern the scaling of tree height and mass. Proc Natl Acad Sci U S A 101:15661-15663.
Nishii I, Miller SM. 2010. Volvox: simple steps to developmental complexity? Curr Opin Plant Biol 13:646-653.
Nishii I, Ogihara S, Kirk DL. 2003. A kinesin, invA, plays an essential role in volvox morphogenesis. Cell 113:743-753.
Nowak MA. 2006. Five rules for the evolution of cooperaiton. Science 314:1560-1563.

Nowak MA, Tarnita CE, Wilson EO. 2010. The evolution of eusociality. Nature 466:1057-1062.
Nozaki H. 1986. Sexual reproduction in *Gonium sociale* (Chlorophta, Volvocales). Phycologia 25:29-35.
Nozaki H, Misawa K, Kajita T, Kato M, Nohara S, Watanabe MM. 2000. Origin and evolution of the colonial volvocales (Chlorophyceae) as inferred from multiple, chloroplast gene sequences. Mol Phylogenet Evol 17:256-268.
Nozaki H, Mori T, Misumi O, Matsunaga S, Kuroiwa H. 2006. Males evolved from the dominant isogametic mating type Curr Biol 16:R1018-R1020.
Nozaki H, Yamada TK, Takahashi F, Matsuzaki R, Nakada T. 2014. New "missing link" genus of the colonial volvocine green algae gives insights into the evolution of oogamy. BMC Evol Biol 14:37.
O'Malley MA, Powell R. 2016. Major problems in evolutionary transitions: how a metabolic perspective can enrich our understanding of macroevolution. Biology and Philosophy 31:159-189.
Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. Bioessays 31:703-714.
Olson BJ, Nedelcu AM. 2016. Co-option during the evolution of multicellular and developmental complexity in the volvocine green algae. Curr Opin Genet Dev 39:107-115.
Olson BJ, Oberholzer M, Li Y, Zones JM, Kohli HS, Bisova K, Fang SC, Meisenhelder J, Hunter T, Umen JG. 2010. Regulation of the Chlamydomonas cell cycle by a stable, chromatin-associated retinoblastoma tumor suppressor complex. Plant Cell 22:3331-3347.
Onmus-Leone F, Hang J, Clifford RJ, Yang Y, Riley MC, Kuschner RA, Waterman PE, Lesho EP. 2013. Enhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping. PLoS One 8:e61762.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061-1067.
Pazour GJ, Agrin N, Lezyk J, Witman GB. 2005. Proteomic analysis of a eukaryotic cilium. Journal of Cell Biology 170:103-113.
Pepper JW, Scott Findlay C, Kassen R, Spencer SL, Maley CC. 2009. Cancer research meets evolutionary biology. Evol Appl 2:62-70.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785-786.
Pines J. 1994. Cell cycle. Ubiquitin with everything. Nature 371:742-743.
Presti JC, Jr., Reuter VE, Cordon-Cardo C, Motzer RJ, Albino AP, Jhanwar SC, Nanus DM. 1996. Expression of the retinoblastoma gene product in renal tumors. Anticancer Res 16:549-556.
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.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86-94.
Queller DC. 2000. Relatedness and the fraternal major transitions. Philos Trans R Soc Lond B Biol Sci 355:1647-1655.
Ramsey G, Pence CH. 2016. Chance in evolution. Chicago: University of Chicago Press.
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.
Rodriguez H, Haring MA, Beck CF. 1999. Molecular characterization of two light-induced, gamete-specific genes from Chlamydomonas reinhardtii that encode hydroxyproline-rich proteins. Mol Gen Genet 261:267-274.
Ropelewski AJ, Nicholas HB, Gonzalez Mendez RR. 2010. MPI-PHYLIP: parallelizing computationally intensive phylogenetic analysis routines for the analysis of large protein families. PLoS One 5:e13999.
Roy RS, Chen KC, Sengupta AM, Schliep A. 2012. SLIQ: simple linear inequalities for efficient contig scaffolding. J Comput Biol 19:1162-1175.
Ruby JG, Bellare P, Derisi JL. 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) 3:865-880.
Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. 2014. BESST--efficient scaffolding of large fragmented assemblies. BMC Bioinformatics 15:281.
Salmela L, Makinen V, Valimaki N, Ylinen J, Ukkonen E. 2011. Fast scaffolding with small independent mixed integer programs. Bioinformatics 27:3259-3265.
Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 22:557-567.
Santorelli LA, Thompson CR, Villegas E, Svetz J, Dinh C, Parikh A, Sucgang R, Kuspa A, Strassmann JE, Queller DC, et al. 2008. Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae. Nature 451:1107-1110.
Sathe S, Durand PM. 2016. Cellular aggregation in *Chlamydomonas* (Chlorophyceae) is chimaeric and depends on traits like cell size and motility. European Journal of Phycology 51:129-138.
Saxer G, Brock DA, Queller DC, Strassmann JE. 2010. Cheating does not explain selective differences at high and low relatedness in a social amoeba. BMC Evol Biol 10:76.
Schmitt R. 2003. Differentiation of germinal and somatic cells in Volvox carteri. Curr Opin Microbiol 6:608-613.
Scott-Phillips TC, Dickins TE, West SA. 2011. Evolutionary Theory and the Ultimate-Proximate Distinction in the Human Behavioral Sciences. Perspect Psychol Sci 6:38-47.

Sebe-Pedros A, Ballare C, Parra-Acero H, Chiva C, Tena JJ, Sabido E, Gomez-Skarmeta JL, Di Croce L, Ruiz-Trillo I. 2016. The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. Cell 165:1224-1237.
Sequeira-Mendes J, Gutierrez C. 2016. Genome architecture: from linear organisation of chromatin to the 3D assembly in the nucleus. Chromosoma 125:455-469.
Sessoms AH, Huskey RJ. 1973. Genetic control of development in Volvox: isolation and characterization of morphogenetic mutants. Proc Natl Acad Sci U S A 70:1335-1338.
Sharpe SC, Eme L, Brown MW, Roger AJ. 2015. Timing the Origins of Multicellular Eukaryotes Through Phylogenomics and Relaxed Molecular Clock Analysis. In: Trillo IR, Nedelcu AM, editors. Evolutionary Transitions to Multicellular Life. Netherlands: Springer. p. 3-29.
Shelton DE, Leslie MP, Michod RE. 2017. Models of cell division initiation in Chlamydomonas: A challenge to the consensus view. J Theor Biol 412:186-197.
Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.
Simpson JT, Durbin R. 2010. Efficient construction of an assembly string graph using the FM-index. Bioinformatics 26:i367-373.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117-1123.
Smith DR, Hamaji T, Olson BJ, Durand PM, Ferris P, Michod RE, Featherston J, Nozaki H, Keeling PJ. 2013. Organelle genome complexity scales positively with organism size in volvocine green algae. Mol Biol Evol 30:793-797.
Smith DR, Hua J, Lee RW. 2010. Evolution of linear mitochondrial DNA in three known lineages of Polytomella. Curr Genet 56:427-438.
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 DR, Lee RW. 2009. The mitochondrial and plastid genomes of Volvox carteri: bloated molecules rich in repetitive DNA. BMC Genomics 10:132.
Smith DR, Lee RW. 2014. A plastid without a genome: evidence from the nonphotosynthetic green algal genus Polytomella. Plant Physiol 164:1812-1819.
Smith J, Queller DC, Strassmann JE. 2014. Fruiting bodies of the social amoeba Dictyostelium discoideum increase spore transport by Drosophila. BMC Evol Biol 14:105.
Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134-1144.

Soderlund C, Nelson W, Shoemaker A, Paterson A. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res 16:1159-1168.
Solari CA, Drescher K, Ganguly S, Kessler JO, Michod RE, Goldstein RE. 2011. Flagellar phenotypic plasticity in volvocalean algae correlates with Peclet number. J R Soc Interface 8:1409-1417.
Solari CA, Drescher K, Goldstein RE. 2011. The Flagellar Photoresponse in Volvox Species (Volvocaceae, Chlorophyceae)1. J Phycol 47:580-583.
Solari CA, Ganguly S, Kessler JO, Michod RE, Goldstein RE. 2006. Multicellularity and the functional interdependence of motility and molecular transport. Proc Natl Acad Sci U S A 103:1353-1358.
Solari CA, Kessler JO, Michod RE. 2006. A hydrodynamics approach to the evolution of multicellularity: flagellar motility and germ-soma differentiation in volvocalean green algae. Am Nat 167:537-554.
Solari CA, Michod RE, Goldstein RE. 2008. Volvox Barberi, the Fastest Swimmer of the Volvocales (Chlorophyceae)(1). J Phycol 44:1395-1398.
Speijer D, Lukes J, Elias M. 2015. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proc Natl Acad Sci U S A 112:8827-8834.
Spudich JL, Sager R. 1980. Regulation of the Chlamydomonas cell cycle by light and dark. J Cell Biol 85:136-145.
Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, et al. 2008. The Trichoplax genome and the nature of placozoans. Nature 454:955-960.
Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, et al. 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466:720-726.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.
Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435-439.
Starr RC. 1968. Cellular differentiation in Volvox. Proc Natl Acad Sci U S A 59:1082-1088.
Starr RC. 1970. Control of differentiation in Volvox. Symp Soc Dev Biol 29:59-100.
Steele RE, Dana CE. 2009. Evolutionary history of the HAP2/GCS1 gene and sexual reproduction in metazoans. PLoS One 4:e7680.
Stein JR. 1966. Growth and Mating of Gonium Pectorale (Volvocales) in Defined Media(1). J Phycol 2:23-28.
Steinback KE, Goodenough UW. 1975. Morphological and Photosynthetic Properties of Digitonin-treated Chloroplast Membranes from the Wild-type and ac-5 Strains of Chlamydomonas reinhardi. Plant Physiol 55:864-869.

Strassmann JE, Queller DC. 2011. Evolution of cooperation and control of cheating in a social microbe. Proc Natl Acad Sci U S A 108 Suppl 2:10855-10862.
Suga H, Chen Z, de Mendoza A, Sebe-Pedros A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sanchez-Pons N, et al. 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325.
Sumper M, Berg E, Wenzl S, Godl K. 1993. How a sex pheromone might act at a concentration below 10(-16) M. EMBO J 12:831-836.
Sumper M, Hallmann A. 1998. Biochemistry of the extracellular matrix of Volvox. Int Rev Cytol 180:51-85.
Szathmáry E. 2015. Toward major evolutionary transitions theory 2.0. Proc Natl Acad Sci U S A 112:10104-10111.
Takebayashi N, Morrell PL. 2001. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. Am J Bot 88:1143-1150.
Tam LW, Kirk DL. 1991a. Identification of cell-type-specific genes of Volvox carteri and characterization of their expression during the asexual life cycle. Dev Biol 145:51-66.
Tam LW, Kirk DL. 1991b. The program for cellular differentiation in Volvox carteri as revealed by molecular analysis of development in a gonidialess/somatic regenerator mutant. Development 112:571-580.
Tam LW, Stamer KA, Kirk DL. 1991. Early and late gene expression programs in developing somatic cells of Volvox carteri. Dev Biol 145:67-76.
Tan EH, Henry IM, Ravi M, Bradnam KR, Mandakova T, Marimuthu MP, Korf I, Lysak MA, Comai L, Chan SW. 2015. Catastrophic chromosomal restructuring during genome elimination in plants. Elife 4.
Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.
Tempel S. 2012. Using and understanding RepeatMasker. Methods Mol Biol 859:29-51.
Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914-939.
Tschochner H, Lottspeich F, Sumper M. 1987. The sexual inducer of Volvox carteri: purification, chemical characterization and identification of its gene. EMBO J 6:2203-2207.
Umen JG. 2011. Evolution of sex and mating loci: an expanded view from Volvocine algae. Curr Opin Microbiol 14:634-641.
Umen JG. 2014. Green Algae and the Origin of Multicellularity in the Plant Kingdom Cold Spring Harb Perspect Biol 6.
Umen JG, Goodenough UW. 2001. Control of cell division by a retinoblastoma protein homolog in Chlamydomonas. Genes Dev 15:1652-1661.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G. 1993. Mitochondrial DNA of Chlamydomonas reinhardtii: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet 24:241-247.
van Gestel J, Tarnita CE. 2017. On the origin of biological construction, with a focus on multicellularity. Proc Natl Acad Sci U S A 114:11018-11026.
Van Valen L. 1973. A new evolutionary law. Evolutionary Theory 1:1-30.
Viamontes GI, Fochtmann LJ, Kirk DL. 1979. Morphogenesis in Volvox: analysis of critical variables. Cell 17:537-550.
Viamontes GI, Kirk DL. 1977. Cell shape changes and the mechanism of inversion in Volvox. J Cell Biol 75:719-730.
Vicedomini R, Vezzi F, Scalabrin S, Arvestad L, Policriti A. 2013. GAM-NGS: genomic assemblies merger for next generation sequencing. BMC Bioinformatics 14 Suppl 7:S6.
Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, Van Heusden P, Singh S, Thevasagayam NM, Prakki SR, Purushothaman K, et al. 2016. Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence Reads and Multi-layered Scaffolding. PLoS Genet 12:e1005954.
Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics.
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.
Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al. 2014. Comprehensive variation discovery in single human genomes. Nat Genet 46:1350-1355.
Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing de novo genome assemblies. Genome Biol 16:207.
West SA, Gardner A, Shuker DM, Reynolds T, Burton-Chellow M, Sykes EM, Guinnee MA, Griffin AS. 2006. Cooperation and the scale of competition in humans. Curr Biol 16:1103-1106.
Whelan NV, Kocot KM, Moroz LL, Halanych KM. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112:5773-5778.
White TD, Lovejoy CO, Asfaw B, Carlson JP, Suwa G. 2014. Neither chimpanzee nor human, *Ardipithecus* reveals the surprising ancestry of both. Proceedings of the National Academy of Science USA 112:4877-4884.
Willensdorfer M. 2009. On the evolution of differentiated multicellularity. Evolution 63:306-323.
Wilson R, Chiang KS. 1977. Temporal programming of chloroplast and cytoplasmic ribosomal RNA transcription in the synchronous cell cycle of Chlamydomonas reinhardtii. J Cell Biol 72:470-481.

Woessner JP, Goodenough UW. 1989. Molecular characterization of a zygote wall protein: an extensin-like molecule in Chlamydomonas reinhardtii. Plant Cell 1:901-911.
Woessner JP, Molendijk AJ, van Egmond P, Klis FM, Goodenough UW, Haring MA. 1994. Domain conservation in several volvocalean cell wall proteins. Plant Mol Biol 26:947-960.
Wolpert L, Szathmary E. 2002. Multicellularity: evolution and the egg. Nature 420:745.
Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.
Wright SI, Kalisz S, Slotte T. 2013. Evolutionary consequences of self-fertilization in plants. Proc Biol Sci 280:20130133.
Yamamoto K, Kawai-Toyooka H, Hamaji T, Tsuchikane Y, Mori T, Takahashi F, Sekimoto H, Ferris PJ, Nozaki H. 2017. Molecular evolutionary analysis of a gender-limited MID ortholog from the homothallic species Volvox africanus with male and monoecious spheroids. PLoS One 12:e0180313.
Yamashita S, Arakaki Y, Kawai-Toyooka H, Noga A, Hirono M, Nozaki H. 2016. Alternative evolution of a spheroidal colony in volvocine algae: developmental analysis of embryogenesis in Astrephomene (Volvocales, Chlorophyta). BMC Evol Biol 16:243.
Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329-342.
Yang L, Koo DH, Li Y, Zhang X, Luan F, Havey MJ, Jiang J, Weng Y. 2012. Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. Plant J 71:895-906.
Zanders SE, Eickbush MT, Yu JS, Kang JW, Fowler KR, Smith GR, Malik HS. 2014. Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. Elife 3:e02630.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.
Zhang J, Li C, Zhou Q, Zhang G. 2015. Improving the ostrich genome assembly using optical mapping data. Gigascience 4:24.
Zhao X, Bramsiepe J, Van Durme M, Komaki S, Prusicki MA, Maruyama D, Forner J, Medzihradszky A, Wijnker E, Harashima H, et al. 2017. RETINOBLASTOMA RELATED1 mediates germline entry in Arabidopsis. Science 356.
Zhu S, Zhu M, Knoll AH, Yin Z, Zhao F, Sun S, Qu Y, Shi M, Liu H. 2016. Decimetre-scale multicellular eukaryotes from the 1.56-billion-year-old Gaoyuzhuang Formation in North China. Nat Commun 7:11500.