

Durham E-Theses

Genomic variation in green macroalgae of the genus Ulva

Jessie Frances Lauze-Cuomo

How to cite:

Lauze-Cuomo, Jessie Frances (2026) Genomic variation in green macroalgae of the genus *Ulva*.
Doctoral thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a <https://etheses.durham.ac.uk/id/eprint/16588/> is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full Durham E-Theses policy](#) for further details.

Abstract

Coastal ecosystems are susceptible to over proliferation of bloom-forming macroalgae due to increases in shoreline development worldwide. One prominent bloom-forming clade, the genus *Ulva*, offers promising model organisms for understanding these environmentally damaging events and plays a developing role in industry as biomass feedstocks for renewable fuel production. This thesis investigates the landscape of genetic variation among cryptic *Ulva* species within this clade. Two novel nuclear genomic assemblies were created from field collected specimens to expand our understanding of species relationships in this genus. Comparative analyses with the first published *Ulva* genome, *U. mutabilis*, indicate a high degree of variability between species, while intraspecific and intergenic comparisons mirror these findings. Proteome size alone is a poor predictor of carbohydrate sulfotransferase levels and cell wall sulfotransferases are elevated in marine chlorophytes as compared to freshwater counterparts. These correlative findings strongly suggest sulfotransferases are an adaptation benefiting stress tolerance in marine macroalgae. Investigation of orthologous genes among more diverse chlorophytes present unique findings of horizontal gene transfer and gene expansion in *Ulva* sulfotransferase activity. A gene family containing sulfotransferases is unique to and expanded in *Ulva* among other chlorophytes. These datasets are rich with data and the prospects for further comparative analysis are strong. Specifically, more direct experimental confirmation of sulfotransferase stress tolerance would add validation to these findings. Further investigation of genetic relationships among *Ulva* species would add context to this developing field.

Genomic variation in green macroalgae of the genus *Ulva*

Jessie Frances Lauze-Cuomo

A thesis presented for the degree of

Doctor of Philosophy



Department of Biosciences

Durham University

United Kingdom

November 2023

Table of Contents

CHAPTER 1: GENERAL INTRODUCTION AND LITERATURE REVIEW.....	13
1.1 CHLOROPHYTE ALGAE	13
1.2 THE CHLOROPHYTE MACROALGAE	17
1.3 ULVA AS A MODEL FOR CHLOROPHYTE MACROALGAE.....	26
1.4 SEAWEED GENOMICS.....	31
1.5 AIMS AND OBJECTIVES.....	31
CHAPTER 2: CREATION OF <i>DE NOVO</i> ASSEMBLIES	42
2.1 INTRODUCTION	42
2.2 METHODS.....	43
2.3 RESULTS.....	50
CHAPTER 3: CHARACTERIZATION OF <i>DE NOVO</i> PSEUDOASSEMBLIES IN THE CONTEXT OF ESTABLISHED REFERENCE GENOMES	61
3.1 INTRODUCTION	61
3.2 METHODS.....	62
3.3 RESULTS.....	63
CHAPTER 4: CHARACTERIZATION OF NEWLY SEQUENCED CHLOROPHYTES	75
4.1 INTRODUCTION	75
4.2 METHODS.....	76
4.3 RESULTS.....	77
CHAPTER 5: ORTHOLOGOUS GROUPS IN SEQUENCED CHLOROPHYTES: EVOLUTION OF GREEN SULFOTRANSFERASES.....	97
5.1 INTRODUCTION	97
5.2 METHODS.....	101

5.3 RESULTS.....	104
CHAPTER 6: GENOMIC VARIATION IN ULVACEAE.....	112
6.1 INTRODUCTION	112
6.2 METHODS.....	113
6.3 RESULTS AND DISCUSSION.....	117
6.4 SUMMARY AND CONCLUSIONS.....	131
CHAPTER 7: FINAL CONCLUSIONS	132
7.1 CREATION OF <i>DE NOVO</i> ASSEMBLIES.....	132
7.2 CHARACTERIZATION OF <i>DE NOVO</i> PSEUDOASSEMBLIES IN THE CONTEXT OF ESTABLISHED REFERENCE GENOMES	133
7.3 CHARACTERIZATION OF NEWLY SEQUENCED CHLOROPHYTES	137
7.4 ORTHOLOGOUS GROUPS IN SEQUENCED CHLOROPHYTES: EVOLUTION OF GREEN SULFOTRANSFERASES.....	141
BIBLIOGRAPHY	145
DIGITAL APPENDIX - TABLE OF CONTENTS	178

List of Figures

Figure 1.1 Microscopic imagery of diverse members of green algae	13
Figure 1.2 Broad view of chlorophyte taxa organization	14
Figure 1.3 Macroalgal diversity distribution along global coastlines	16
Figure 1.4 Cell type and arrangement of select individuals of varying body plans in the class Ulvophyceae	18
Figure 1.5 Organization of the core chlorophytes overlaid with cellular organization pattern.....	19
Figure 1.6 Phylogeny of core chlorophytes overlaid with cellular wall structural components.....	20
Figure 1.7 Fouling of transportation vessels and release of ballast water may accelerate the spread of invasive species globally	21
Figure 1.8 Green tide global events.....	22
Figure 1.9 Visualization of the two distinct morphological forms of <i>Ulva</i>	26
Figure 1.10 Simplified version of the 2018 <i>Ulva</i> phylogeny	28
Figure 1.11 Life cycle of <i>Ulva</i>	30
Figure 1.12 Algal genome comparison of relative genome size and number of genes between three core chlorophyte groups.....	33
Figure 1.13 Collinearity analysis chord diagram with <i>U. prolifera</i> and <i>U. mutabilis</i> showing areas of synteny	34
Figure 1.14 <i>Caulerpa lentillifera</i> morphology	35
Figure 1.15 Phylogenetic tree of <i>Caulerpa</i> relation to other chlorophytes.....	36
Figure 1.16 Divergence time calculated from time-calibrated phylogeny	37
Figure 1.17 Simplified model indicating additions of “genomic novelty”	38
Figure 2.1 Sample collection from two locations sites in the United Kingdom.....	44

Figure 2.2 Culture sample in preparation for DNA extraction.....	45
Figure 2.3 Model depicting SNP identification.....	47
Figure 2.4 Model depicting resulting sequencing extraction	47
Figure 2.5 Possible annotation detection pathways via the EuGene platform	49
Figure 2.6 Read workflow for Ward files	50
Figure 2.7 Read workflow for Langston files	52
Figure 2.8 Bandage plots for Ward and Langston.....	56
Figure 2.9 Control analysis dot plots of Ward and Langston.....	57
Figure 2.10 Contig length frequency distribution	58
Figure 2.11 Ward contig 42 annotated alongside <i>A. thaliana</i>	59
Figure 2.12 Langston contig 619 annotated alongside <i>A. thaliana</i>	60
Figure 3.1 <i>Ulva mutabilis</i> bandage plot	64
Figure 3.2 Dot plot analysis of query Ward and subject <i>U. mutabilis</i> genomes	66
Figure 3.3 Dot plot analysis of query Langston and subject <i>U. mutabilis</i> genomes...	68
Figure 3.4 Dot plot data extraction of identity matches between each query and the target, <i>U. mutabilis</i>	70
Figure 4.1 <i>Caulerpa lentillifera</i> bandage plot.....	79
Figure 4.2 <i>Ulva prolifera</i> bandage plot.....	79
Figure 4.3 Dot plot analysis of query <i>C. lentillifera</i> and subject <i>U. mutabilis</i> genomes	82
Figure 4.4 Dot plot analysis of query <i>U. prolifera</i> and subject <i>U. mutabilis</i> genomes	84
Figure 4.5 Coordinates of low to mid-range variation hits from <i>U. mutabilis</i> and <i>U.</i> <i>prolifera</i> dot plot.....	88
Figure 4.6 Comparison of open reading frame count between target and query	90

Figure 4.7 Comparison of chloroplast genomes.....	94
Figure 5.1 Phylogeny of four key green algal species based on proteome data.....	97
Figure 5.2 Catalytic reaction with sulfotransferases.....	98
Figure 5.3 FTIR spectroscopy of Ulva cell wall	100
Figure 5.4 Extraction protocol for cell wall fractions	102
Figure 5.5 Phylogeny of proteomes used in Orthofinder analysis	104
Figure 5.6 Number of carbohydrate SOT as a function of proteome size.....	106
Figure 5.7 Species phylogeny resulting from the orthogroup SOT analysis.....	108
Figure 5.8 Maximum likelihood phylogenetic tree of bacterial and green/red algal sequences.....	109
Figure 5.9 Maximum likelihood tree showing rearrangement based on changes of function parameters	110
Figure 5.10 FTIR spectroscopy of Ulva and Caulerpa.....	111
Figure 6.1 Mapped locations of samples compiled in the global collection.	125
Figure 6.2 Mapped locations of samples in RADseq analysis	126
Figure 6.3 Match length and percent identity of gene hits which returned significant E-values	129
Figure 6.4 Relationship between total read count and aligned read count on logarithmic and simple numerical scales	130

List of Tables

Table 2.1 Basic statistics of de novo pseudoassemblies, Ward and Langston, compared to <i>U. mutabilis</i>	54
Table 2.2 Contigs from Ward and Langston pseudoassemblies with base length greater than 1,000.	55
Table 3.1 Percent variation accounted for along <i>U. mutabilis</i> contigs by pseudoassembly.	71
Table 3.2 Annotation of strongest query hits between pseudoassemblies and <i>U. mutabilis</i> reference genome.	72
Table 3.3 Ward and Langston BLAST results from non-matched <i>mutabilis</i> hits	73
Table 4.1 Summary statistics and basic demographics of sequenced Chlorophytes..	78
Table 4.2 High identity matches with reference genome indicate conservation of genomic structure.	86
Table 4.3 Open reading frame count from low identity dot plot matches along the central diagonal	91
Table 4.4 Basic statistics of <i>U. mutabilis</i> and <i>U. prolifera</i> chloroplast genomes	93
Table 5.1 Comparison of cell wall components and functional groups across major green algal groups	99
Table 5.2 Proteome data and cell wall components of red and green micro- and macroalgae.	105
Table 5.3 Species involved in the Orthofinder analysis and their respective contributions to the total genes assigned and orthogroup count	107
Table 6.1 Data fields sought for each sample in the crude collection.	114
Table 6.2 Standards and criteria to be met for inclusion in sample database	114
Table 6.3 Categorical organization of sample types	115

Table 6.4 DNeasy protocol modifications for tissue-specific extractions	116
Table 6.5 Description of protocol modifications for dried sample extractions and corresponding test results	117
Table 6.6 Presentation of unique specimens in the global collection in order of ascending northern latitude	119
Table 6.7 Database identifications of sequences from aligned RADseq files.....	127

List of abbreviations

The following table describes abbreviated terms used throughout this document and their significance in the thesis. The page on which each example is first used and defined is also included.

Abbreviation	Meaning	Page
LSU	Large subunit	27
<i>U. mutabilis</i>	<i>Ulva mutabilis</i> , <i>Ulva compressa</i>	31
NGS	Next generation sequencing	32
PAF	Pairwise mapping format	67
LTR	Long terminal repeat	80
LIE	Long interspersed element	80
Ulvmu, Uco	<i>Ulva mutabilis</i> , <i>Ulva compressa</i>	86
SSU	Small subunit	87
RMSD	root mean square deviation	90
Upro	<i>Ulva prolifera</i>	91
SOT	sulfotransferase	97
PAPS	3'-Phosphoadenosine 5'-Phosphosulfate synthase	98
FTIR	Fourier-transform infrared	99
MYA	Million years ago	100
HGT	Horizontal gene transfer	101
IR/NIR	Infrared/near infrared	103
CHST	Carbohydrate sulfotransferase	104
GBS	Genotyping by sequencing	112
RADseq	Restriction-site associated sequencing	113
WGS	Whole genome sequencing	113
CTAB	Cetyltrimethylammonium bromide	136

Declaration

This thesis is based on research carried out at the Durham University Department of Biosciences between October 2016 to December 2020. No part of this thesis has been submitted elsewhere for any other degree or qualification. This work and written thesis is in the words of the author.

Statement of Copyright

The copyright of this thesis rests with the author. No quotation from it should be published without the author's prior written consent and information derived from it should be acknowledged.

Acknowledgements

Thank you to my advisor, Dr John Bothwell, for investing time and effort into my development as a scientist. Thank you for welcoming me into your family while I was so far away from my own. I will always look back fondly on Thanksgiving meals in the UK. Thank you for helping Sammy and I along the way; for this, we are so fortunate.

Thank you to my friends and family. To my mother, brother, in-laws, and closest friends: I would not have made it without you. Your video calls, care packages, and visits kept me motivated while an ocean separated us. Thank you to my sister-in-laws for bringing the newest loves of my life into this world, Emilia, Josephine, John, and Cameron. To the friends I met along the way, I am so lucky to know you.

Thank you to my other half, Anthony, who became my husband during this studentship. Thank you for encouraging me to take risks, invest in myself, and be adventurous. You believed in me, even when I didn't. Before this chapter was finished, we experienced the highest of highs and lowest of lows. It is all inevitably a part of our story, and what a wonderful story it is. Angelo, Leo, and I are so lucky to have you.

Thank you, Leo, my light in the darkness. You saved me. Mommy loves you endlessly and without measure.

And finally, thank you to my travel companion, Sammy, who fought against all odds to stay with me to the very end. There does not exist a more loyal and loving companion. You are so missed.

Chapter 1: GENERAL INTRODUCTION AND LITERATURE REVIEW

1.1 Chlorophyte algae

Members of the phylum Chlorophyta are the lesser-studied cousins of land plants. The most widely familiar chlorophytes are perhaps the green seaweeds and freshwater phytoplankton, although the lineage is diverse and includes many soil-dwelling and extremophile species (Lewis 2007).

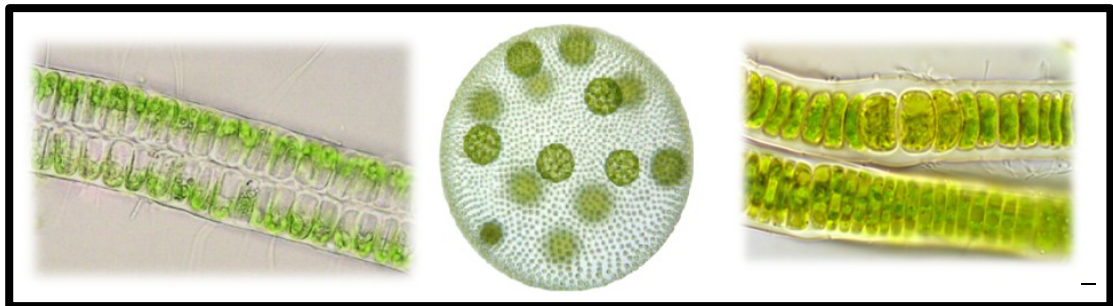


Figure 1.1 Microscopic imagery of diverse members of green algae, specifically (left) brackish and saline alga Ulva, (center) colonial freshwater alga volvox, and (right) drought tolerant and semi-terrestrial prasiola (accessed via Algaebase 2021) (scale is approximately 10 μm).

The green algae display many body plans and life history strategies. Living in freshwater (mainly volvocales), saltwater (eg. Mamelliales and ulvophytes), or even terrestrially (many trebouxioophytes), they may be unicellular (micro), multicellular (macro), or colonial, with mono- or multinucleate cells (Figure 1.1) (Domozych et al. 2016, Beardall et al. 2008, Holzinger et al. 2017). Their reproduction strategy also varies, with some species undergoing asexual reproduction by fission, budding, or fragmentation with motile zoospores, while other undergo sexual reproduction, and with many species employing more than one reproductive strategy (De Clerck 2018).

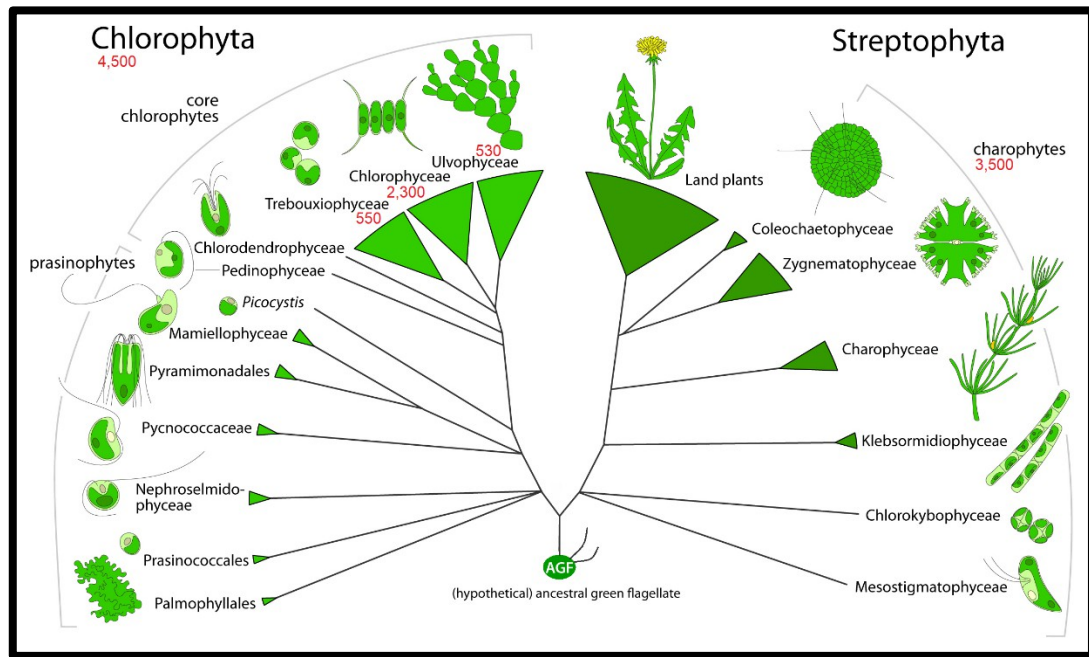


Figure 1.2 Broad view of chlorophyte taxa organization with relevant approximate species counts in red (adapted from Leliaert et al. 2012 [updated 2013] and Guiry 2012).

Defining taxa within the Chlorophyta proves complicated and remains non-static, with a current estimate of 4,500 species members (Figure 1.2) (Gulbrandsen et al. 2021). This diverse set of organisms once played a critical role in the succession of land-based plants and co-evolving fauna, and at present, global narratives of invasive species, coastal pollution, and sustainable carbon sources are not possible without its inclusion (Del Cortona 2020).

1.1.1 Evolution of multicellularity

Evolutionarily, the chlorophytes represent half of the split that occurred between 970 and 1244 million years ago, which separated them from the streptophyte lineage that later evolved into today's land plants (Wang et al. 2020). Separate endosymbioses of chlorophyte algae gave rise to chloroplasts in the euglenids of the Excavata and the chlorarachniophytes of the Rhizaria (Gibbs 1978, Rogers et al. 2007).

The developmental progression from single celled eukaryotes to complex multicellular ones is rare, but evolutionarily important. Complex multicellular life forms have evolved in both animal and plant clades (Rensing 2016, Sebe-Pedros et al. 2017, Cock et al. 2010) and chlorophyte algae represent an important group of research organisms with which to study this process.

1.1.2 Environmental importance of the chlorophytes

According to our best current estimates, chlorophytes are responsible for 5-10% of global primary production. The total contribution of the chlorophyte lineage to oceanic net primary production has been estimated at 7 PgC.y⁻¹ from remote sensing data, with particular richness in coastal and equatorial waters (Rousseaux and Gregg 2014). Estimates of their contribution to global freshwater primary production are less common, but global wetlands contribute approximately 4 PgC.y⁻¹ (Huston and Wolverton 2009). Chlorophyte plankton of the Trebouxiophyceae or Chlorophyceae dominate in many parts of the phytoplankton succession in these freshwater ecosystems across a range of trophic states (Reynolds et al. 2002, Wang et al. 2018). This suggests that chlorophytes perhaps contribute another 2 PgC.y⁻¹, for an estimated chlorophyte net primary production of approximately 9 PgC.y⁻¹, or about 10% of the global total of 100-110 PgC.y⁻¹ (Field et al. 1998). For comparison, this amounts to slightly more than the global contribution of either the World's cultivated croplands or the temperate forests, both of which have been estimated between 6-7 PgC.y⁻¹ (Field et al. 1998, Huston and Wolverton 2009).

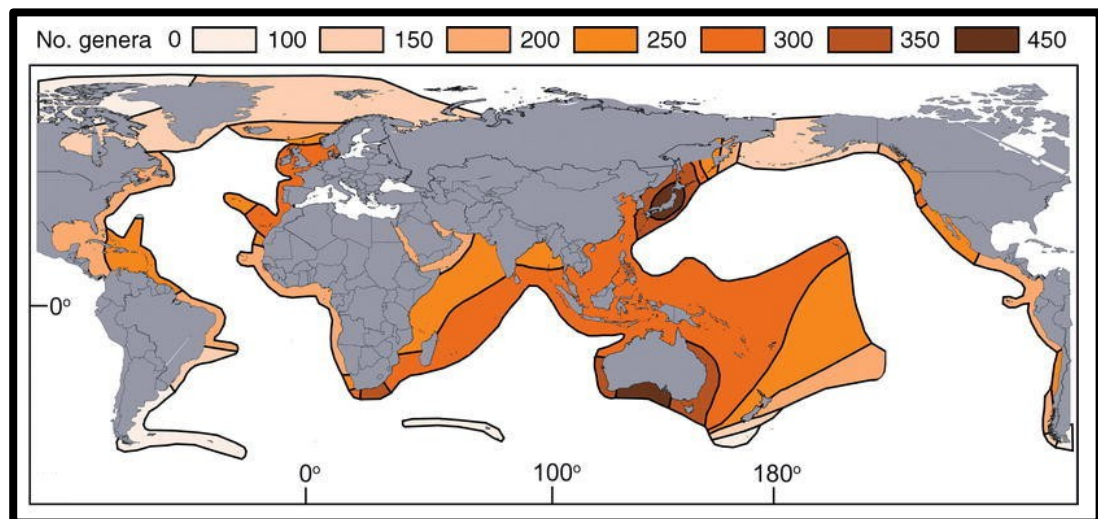


Figure 1.3 Macroalgal diversity distribution along global coastlines. Saturation of orange/brown hue indicates species richness, in terms of genera count. Diversity is highest in temperate and tropical regions, lowering towards the poles (adapted from Kerswell 2006).

An elegant 2018 report of microalgae composition among worldwide coastal waters surveyed 141 marine locations ‘quasi-simultaneously’ in an event called The Ocean Sampling Day (Tragin and Vaultot 2018). The survey found chlorophyte composition constituted between 10 and 100% of site-specific phytoplanktonic presence, with the global estimate around one third. A macroalgal equivalent study of this magnitude does not yet exist, likely due to what would be a far more complex and exhaustive sampling and identification effort. However, measures of algal diversity along global coastlines are rich in tropical and temperate latitudes (Figure 1.3) and are mimicked in a similar distribution of Bryopsidales (Kerswell et al. 2006). Importantly, Pacific coastlines were not included in this study, although these areas are known to be diverse and productive ecosystems (Bittick et al. 2019, Fernández-García 2011, Rincón-Díaz et al. 2021, Dewinta et al. 2021, Santelices and Marquet 2003).

1.1.3 *Economic importance*

The chlorophytes are less important economically, although some of their constituent microalgal species are grown at commercial scales, most notably *Chlorella* sp. (Trebouxiophyceae) for animal feed and *Haematococcus* sp. (Chlorophyceae) for the pigment, astaxanthin (Kotrba et al. 2015, Bubrick 1991). Proof-of-concept work has demonstrated that chlorophyte microalgae (Miao et al. 2006, Scragg et al. 2003) and seaweeds (Matsui and Koike 2010, Trivedi et al. 2016) may each have bioenergy potential, but the economic viability of algal biofuels remains an open question (Kenny and Flynn 2017, Shuba and Kifle 2018).

More traditionally, chlorophyte seaweeds are well-established delicacies in many countries, most notably *Ulva* sp., *Codium* sp., and *Caulerpa* sp. as ‘Limu’ in Hawaii (Abbott 1984) and *Monostroma* sp. and *Ulva* sp. as ‘Aonori’ in Japan (Chapman 1950, Tanaka et al. 2020). Perhaps the most important economic relevance of the chlorophytes are the ‘green tides’ (Smetacek and Zingone 2013). These blooms show significant annual variability, but a large bloom can deposit over one million tons of biomass along a few miles of shoreline, costing several million dollars in losses and removal (D. Liu et al. 2013).

1.2 The chlorophyte macroalgae

Chlorophyte algae group into microscopic and macroscopic forms. Perhaps counterintuitively, not all macroalgae are multicellular (Figure 1.4). Within ulvophytes alone there are multiple cellular arrangements and nuclear patterns, many of which are coenocytic and multinucleate.

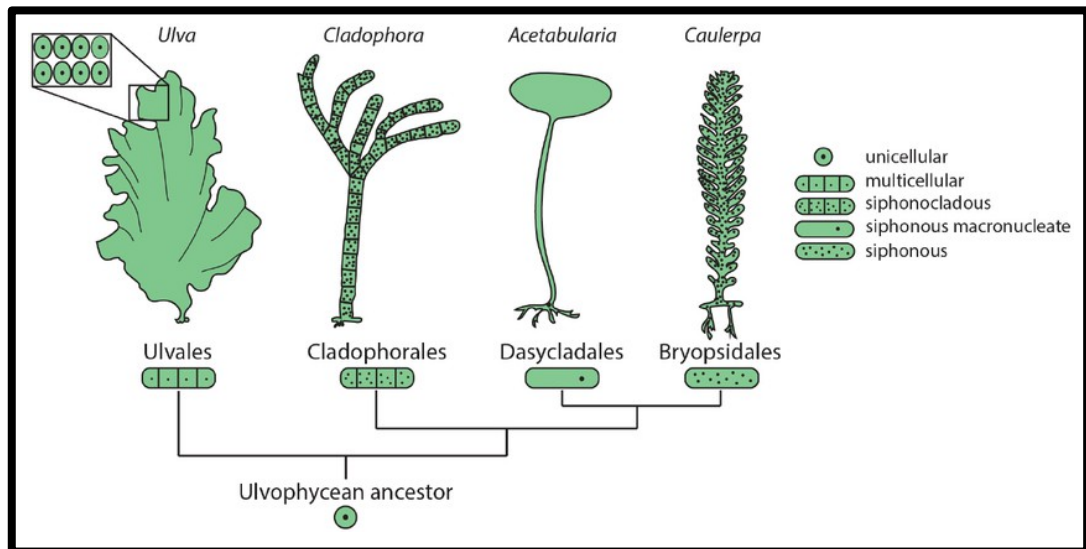


Figure 1.4 Cell type and arrangement of selected individuals of varying body plans in the class Ulvophyceae (adapted from Coneva and Chitwood 2015).

Green macroalgae present similar cellular features as other eukaryotic photosynthesizers, including land plants. Chlorophyll a and b are the primary photosynthetic pigments, with xanthophylls and carotenes as accessory pigments. The primary storage of carbohydrates is in the form of starches. Even the primary land plant reproductive strategy, ‘alternation of generations’, is founded within green macroalgae, although other strategies also exist.

Nonetheless, within the core Chlorophyta alone, there is an astonishing amount of structural, ecological, and developmental diversity. The core chlorophytes are classified into three classes, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae, often referred to as “UTC” (Figure 1.5). These organisms evolved from early diverging prasinophytes (Leliaert et al. 2012). Chlorophyceae and Trebouxiophyceae are mainly dominant in terrestrial and freshwater ecosystems, while Ulvophyceae evolved largely in coastal marine environments.

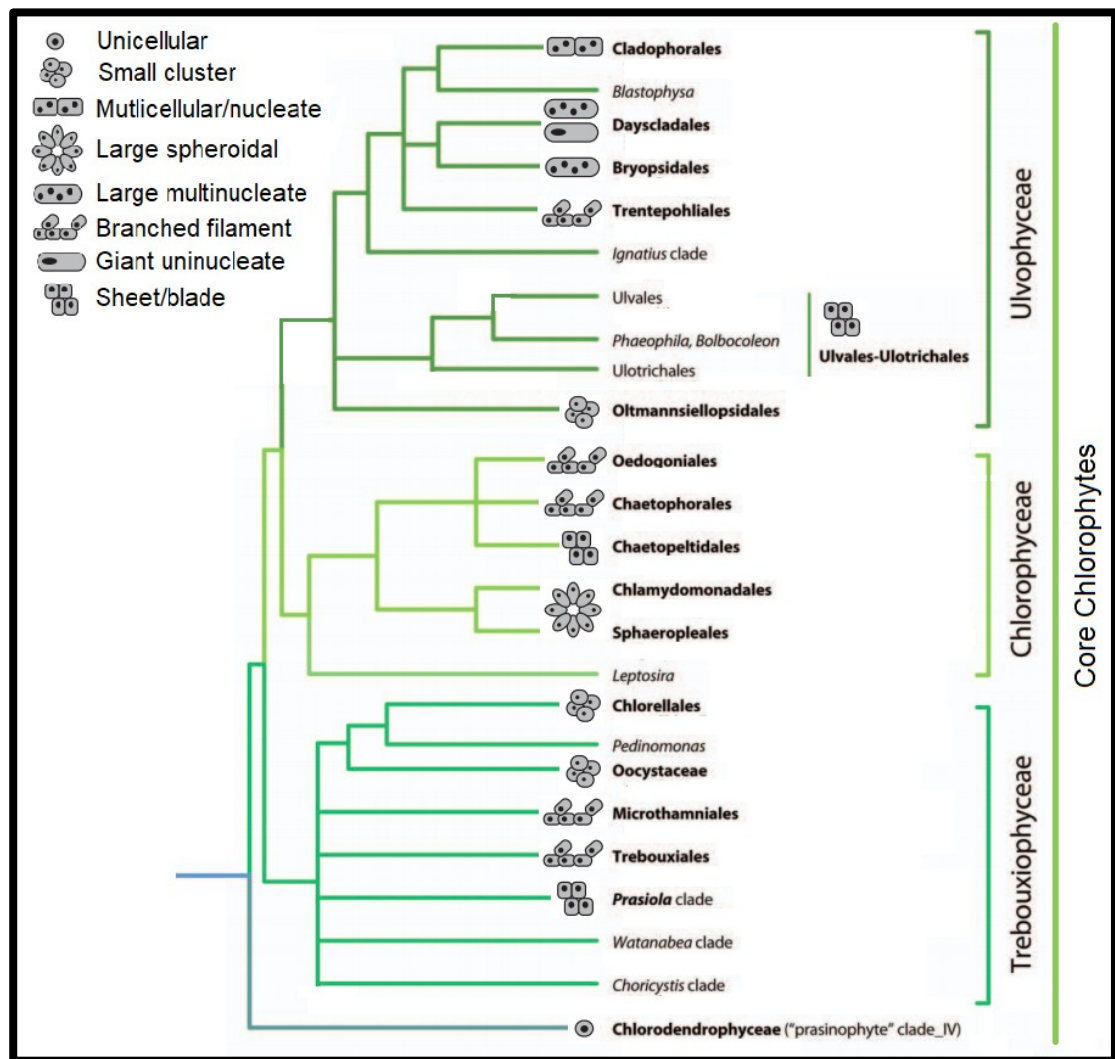


Figure 1.5 Organization of the core chlorophytes overlaid with cellular organization pattern for selected taxa (adapted from Leliaert et al. 2012 and Umen 2014).

These delineations extend to reproductive strategies. Members of the Chlorophyceae and Trebouxiophyceae reproduce sexually in haplontic fashion, where the only diploid form is the zygote, and may also reproduce asexually via cell division, zoospores, or autospores (Leliaert et al. 2012). Members of Ulvophyceae follow a haplodiplontic sexual reproduction mode, in which there are two vegetative forms that cycle in ploidy. Uniquely, ulvophyte cellular division does not involve phycoplasts, which are groupings of microtubules that orient perpendicular to the spindle axis and aid in the formation of the cell wall that divide daughter cells in the other chlorophytes.

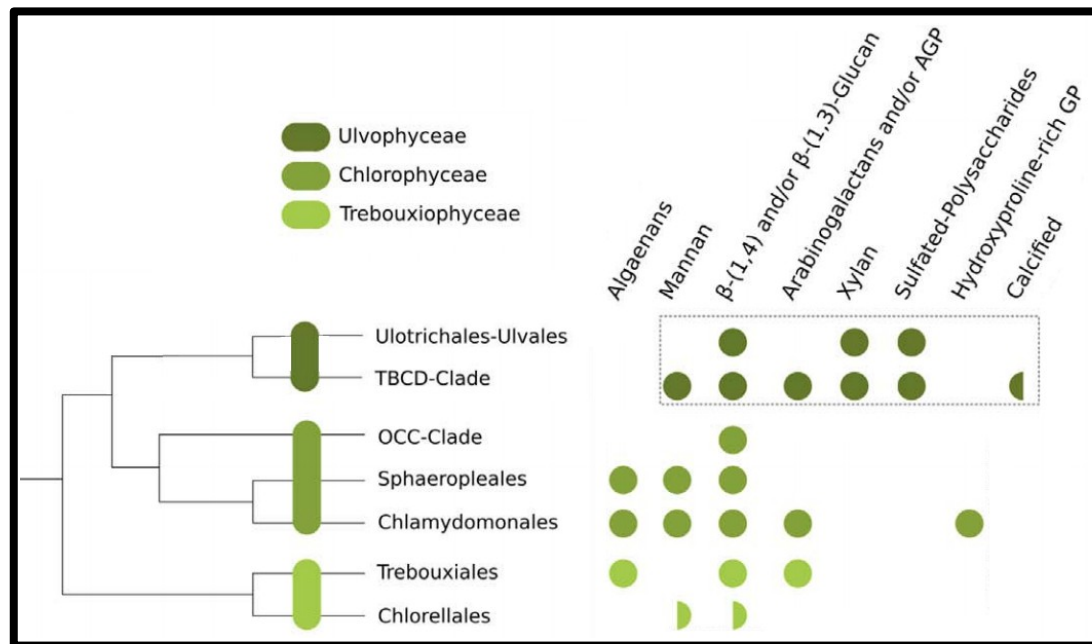


Figure 1.6 Phylogeny of core chlorophytes overlaid with cellular wall structural components. Circles indicate majority/proportion of group with the specified trait. All taxa are characterized by a multi-layer, polysaccharide cell wall. The boxed examples are entirely macroalgal representatives (adapted from Baudelet et al. 2017).

Cell wall composition is another area of great diversity among core chlorophytes (Figure 1.6). Ulvophyte cell walls, for example, are rich in polysaccharides, principally cellulose and a class of sulfated polysaccharide appropriately called ulvans and may also include β -mannans and -xylans (Ciancia et al. 2012, Estevez et al. 2009, Percival 1979, Lahaye and Robic 2007). Chlorophyte and trebouxioophyte cell walls also contain cellulose, but may also incorporate crystalline glycoproteins or algaenan, a polymer that may prevent against viral, bacterial, or fungal infection (Kirk 1998, Voigt et al. 2001, 2007, Cordeiro et al. 2006, Estevez et al. 2008, Rodrigues and da Silva Bon 2011, Domozych et al. 2012, Dunker and Wilhelm 2018).

1.2.1 Ecological importance of the chlorophyte seaweeds

While *Ulva* can contribute to an environmentally healthy ecosystem through primary production and sequestration of harmful pollutants, it has the potential to rapidly overgrow, leading to eutrophication events. During such events, effluent containing

inorganic, growth-limiting nutrients are carried into coastal waterways where they promote unnatural levels of macroalgal growth (Fletcher 1996). *Ulva* is unusually fast growing and makes efficient use of excess dissolved nitrogen. So close is this association that growth rates can be predicted by annual dissolved nitrogen concentration in discrete locations (Teichberg et al. 2010). Nitrogen and phosphorus, common in wastewater and fertilizer runoff, are therefore believed to be the limiting nutrients that, when present in excess, drive rapid proliferation in *Ulva* (Lyngby et al. 1999, Villares et al. 1999).

Algal blooms create instability in coastal ecosystems, as macroalgae over-proliferate in relatively small areas. Fast-growing *Ulva*, or other invasives such as *Caulerpa* sp., outcompete resident macroalgae, either by nutrient uptake or reduction in space and light



Figure 1.7 Fouling of transportation vessels and release of ballast water may accelerate the spread of invasive species globally (Bax et al. 2003, Rosenhahn et al. 2008).

availability (Camargo and Alonso 2006). The increasing connectivity of the oceans via natural or anthropogenic avenues means that these species are rapidly circulated around the globe (Figure 1.7). In the later stages of eutrophication, sedimentation and bacterial decomposition can create hypoxic conditions, which often result in a loss of species diversity (flora and fauna) from large scale die-offs on a seasonal basis (Diaz and Rosenberg 1995).

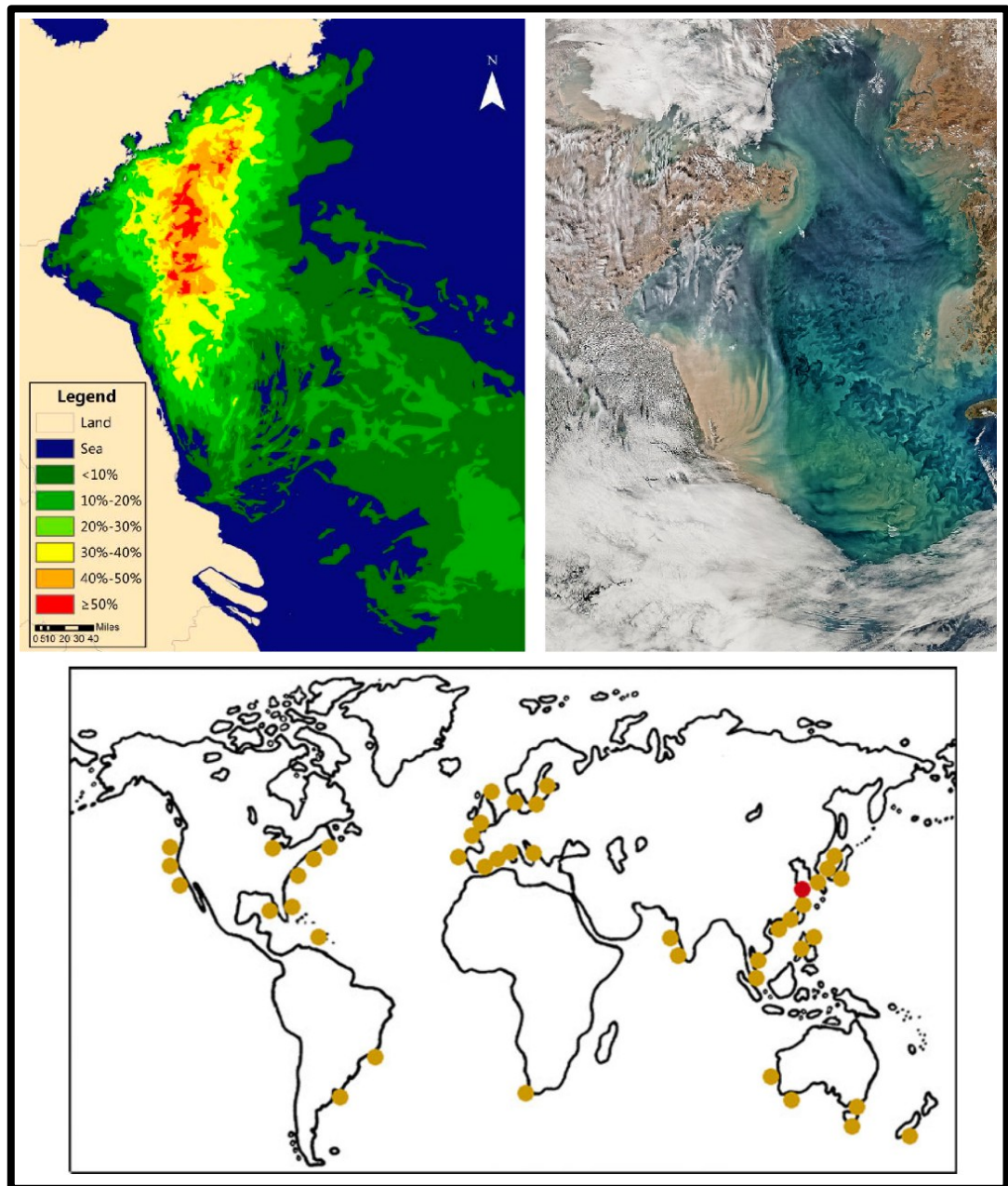


Figure 1.8 Green tide global events, especially in the Yellow Sea, China: (top left) probability distribution of *U. prolifera* between 2016-2018, (top right) satellite imagery of Yellow Sea with algal biomass visible from space in the same location, and (bottom) global distribution of documented harmful algal blooms worldwide. Red point indicates location of *Ulva* bloom in the Yellow Sea, China (adapted from Cao et al. 2019 and Ye et al. 2011).

The 2008 Beijing Olympics remains perhaps the best-known example of the immense scale of these eutrophic events. In just two months over 1 million tons of *Ulva* had grown in the Yellow Sea covering an area of 400 km² (Liu et al. 2013, Zhang et al. 2014, Hu et al. 2017) (Figure 1.8). In this particular instance, accelerated *Ulva* growth was attributed, in part, to effluent from nearby farmland and aquaculture (Liu et al.

2013). Such ‘green tide’ eutrophic events are widespread, rapid, and are likely happen globally with increasing regularity as a result of increased agricultural land usage, climate warming, and population flux in coastal cities and towns. It is important to note that these events are not occurring in isolated, low population density areas, but rather the opposite, and on a scale visible from space, with annual blooms in the Gulf of Mexico, English Channel, and North American Great Lakes (Smetacek and Zingone 2013, Racine et al. 2021, Lefebvre and Devreker 2020, Garnier et al. 2019, Gladyshev and Gubelit 2019).

1.2.2 *Economic importance of the chlorophyte seaweeds*

Beyond their basic evolutionary and ecological interests, *Ulva* species also have practical, economic applications. Given the increasing prevalence of aquaculture worldwide, algal-based feed products are possible steps toward an environmentally conscious economy. A collaborative report produced by the World Bank, the Food and Agriculture Organization of the United Nations, and the International Food Policy Research Institute predicts that over half of world seafood demands will be supplied by aquaculture by 2030 (World Bank 2013). There is, therefore, an obvious market for commercial-scale algal growth and one that is notably underrepresented by green macroalgae (Moreira et al. 2021).

In fact, the rapid growth of *Ulva* spp. make them particularly attractive species for aquaculture, as well as their candidacy for biofuel production. Moving to fuel systems based on biological tissue is environmentally attractive, since biomass is renewable, unlike the planet’s depleting fossil fuel reserves. Global reliability on fossil fuels as a primary energy source has, of course, also led to accumulation of heat-trapping gas species, most notably carbon dioxide (Garret et al. 1992, Wuebbles and Jain 2001, Montzka et al. 2011). Decades of this activity are manifesting as increased global temperatures, elevation of sea levels worldwide, and ocean acidification, all of which

drastically alter the state and livability of ocean habitats (Sahoo et al. 2012). Marine and brackish macroalgae, such as *Ulva*, represent a potential source of renewable biomass because they are rich in carbohydrates and grow rapidly, increasing yield in the production of fuels such as methanol and ethanol without competing for space and resources with land-based agricultural crops (Dang et al. 2012, Demirbas 2010).

Not only are seaweeds a potential candidate for emerging large-scale biofuel production efforts, but their biological demands include uptake of carbon dioxide (Sahoo et al. 2012). *Ulva* sp. may employ two different modes of carbon uptake given the conditions of their environment, making them adaptable and productive in a range of environments (Beer et al. 1990). Thus, atmospheric carbon dioxide can be assimilated into the tissues of these seaweeds and removed from the atmosphere and surface waters, where they are most environmentally problematic (Ortega et al. 2019). However, while many studies recognize that *Ulva* has potential as a biofuel source, it is generally agreed that more work must be done on optimizing the transformation of the raw material into the final product to ensure scalability and cost-efficiency (Milledge et al. 2014, Khan et al. 2016, Bruhn et al. 2011). This process may be enabled by expanding our current knowledge of the biology of these organisms.

Currently, *Ulva* has a presence in commercial consumer industries including food, medicine, and biofertilization (Hong et al. 2007, Zodape 2001, Chbani et al. 2015). Seaweeds are rich in many crucial metals, in particular calcium, and the biochemical

composition of *Ulva* (high in protein, amino acids, and soluble fiber and low in fat) have made them the focus of therapeutic studies to alleviate symptoms of metabolic diseases while delivering fewer side effects than traditional medicinal supplements (Fleurence 1999, Mohamed et al. 2012). Polysaccharide extract from *Ulva* has even been shown to support hepatic health in model species and reduce hypertension, although much work remains to be done in this field (Sathivel et al. 2008).

Humans are not the only ones who benefit from the hearty nutrition of *Ulva*. In commercial aquaculture, efficiency in animal stock growth is a critical consideration. Inclusion of *Ulva* in the diet of cultured fish, such as bass and tilapia, has been shown to support and improve fish growth and yield favorable body composition (Valente et al. 2006, Ergün et al. 2009). In a co-culturing experiment, the use of *Ulva* as shrimp feed was more efficient than traditional formulas, with increased feed-biomass conversion ratios and increased growth rates (Cruz-Suárez et al. 2010). Even ruminant livestock may benefit from *Ulva* as a substituent in their regular diet as a high protein food supplement (Arieli et al. 1993, Machado et al. 2015).

Concurrently, the bioaccumulative properties of macroalgae have been investigated for their use in remediation of contaminated water systems. As it is a ubiquitous coastal species, this potential is increasingly examined in *Ulva*. Coastal marine environments are often subject to anthropogenic effluents, exposing members of these habitats to hazardous chemicals at unnatural levels. Bioremediation is the use of biological material to combat environmental contamination, and *Ulva* has been utilized in clean-up efforts of this kind because of its absorptive properties and prodigious growth. Studies have displayed the capability of *Ulva* to efficiently remove hazardous metals in suspension, such as copper and mercury, making them effective bioremediators (Vijayaraghavan et al. 2004, Zeroual et al. 2003). This capability may be particularly

advantageous in waterways alongside industrialized communities where effluent is often metal-contaminant rich (Roberts et al. 2015).

1.3 *Ulva* as a model for chlorophyte macroalgae

1.3.1 *Ulva* diversity

Species within the genus *Ulva* are well-known for being difficult to distinguish. Morphological identifiers between species are subtle and vary, even among conspecifics (Blomster et al. 1998, Hayden and Waaland 2004). Correct species identification is complicated by morphological plasticity in response to physical cues, such as eutrophic conditions or salinity. These conditions can induce a distromatic (two layered) sheet or a tubular phenotype, for example (Blomster et al. 2002, Blomster et al. 1998) (Figure 1.9). As an example of this complexity, *Ulva* and

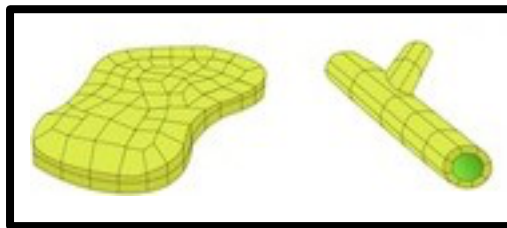


Figure 1.9 Visualization of the two distinct morphological forms of *Ulva* spp.: (left) the distromatic sheet and (right) the cylindrical tube.

Enteromorpha were once grouped as a single genus characterized by double layer branching or thalli. Later, they were separated into two genera based on blade or tubular forms, only for this distinction to be retracted around twenty years ago on the basis of

molecular data (Hayden et al. 2003). Other abiotic factors that can affect morphology include temperature, season, and wave exposure. Individuals may also exist adhered to rocky substratum via holdfast or as free-floating individuals (Loughnane et al. 2008).

On this basis, morphology is increasingly unreliable as the appropriate factor on which to build genus-level phylogenies, regardless of the fact that efforts have been extended to characterize species by unique morphological signatures for decades (Bliding 1968, Koemann & van den Hoek 1981, Hoeksema & van den Hoek 1983, Tanner 1986, Dion et al. 1998). Molecular-based phylogenies are recognized as a necessity for developing *Ulva* phylogenetic research. We are beginning to understand the shortcomings of this approach, and it is likely that future work in the field will center around redefining historical records and species definitions with the inclusion of molecular-marker IDs and mitochondrial and/or chloroplast genome sequencing (Fort et al. 2021, Tran et al. 2022). Two DNA-based methods in particular have risen to prominence: nuclear ribosomal RNA ITS and 5.8S sequencing and DNA-based *rbcL* gene sequencing.

Nuclear rRNA sequencing involves whole genome DNA extraction, followed by polymerase chain reaction (PCR) using primers for the appropriate target: the non-functional Internal Transcribed Spacer (generally ITS 1 or 2) or the noncoding 5.8S RNA large subunit (LSU) component of the ribosome. The products are then sequenced and compared (Tan et al. 1999, Hiraoka et al. 2011).

Despite the utilities of ITS and 5.8S sequences, the *rbcL* (large subunit ribulose biphosphate carboxylase/oxygenase) gene sequence has emerged as the more popular and consistent tool in this field. Its reliability has been tested in species identification over the past decade, and it is the most widely utilized method for this purpose in continued *Ulva* studies (Hayden and Waaland 2002, Hayden and Waaland 2004, Loughnane et al. 2008, Guidone et al. 2013). Notably, some have proposed *tufa*, which encodes the chloroplast elongation factor Tu, as the most successful DNA-based identification target in terms of species delimitation across analytical methods (Weinberger et al. 2021), but it does not seem to be the most widely applied

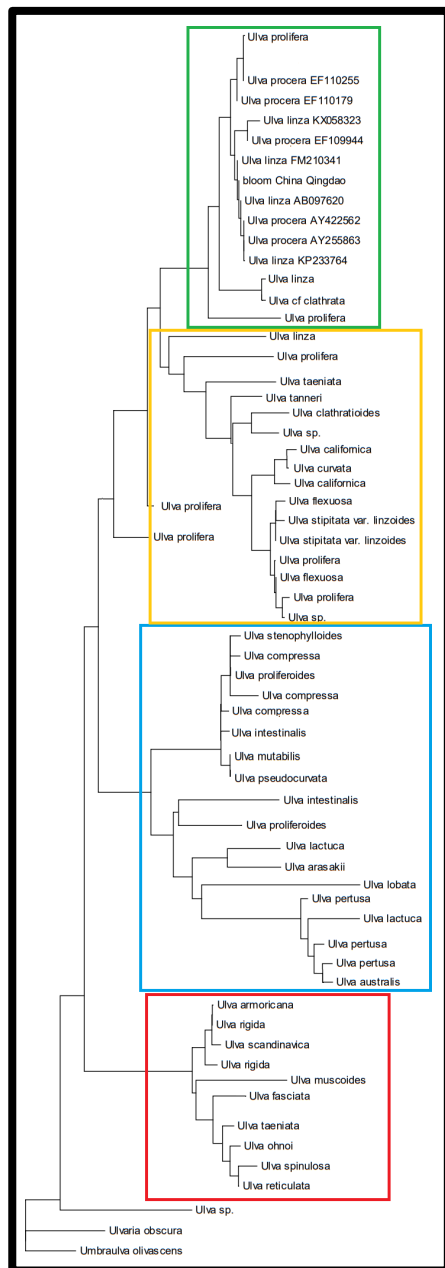


Figure 1.10 Simplified version of the 2018 *Ulva* phylogeny indicating four unique clades: (green) *linza/prolifera*, (yellow) *intestinalis/lactuca*, (blue) *pseudocurvata/compressa*, and (red) *rigida/ohnoi* (modified from De Clerck et al. 2018).

method in the literature, perhaps due to its relative novelty (Vieira et al. 2016, Tran et al. 2022).

Indeed, often both ITS and *rbcL* sequences are of interest to resolve phylogenies in these studies. These kinds of ‘scanning’ methods, in which a short, unique sequence is sought from the pool of total possible sequences, are the basis of DNA-barcoding. While the approach was presented by Hebert et al. in 2003, its practical use among model research organisms is a more novel technology, with chloroplast whole genome sequencing serving as another popular approach in these comparative analyses (Mahendran and Saravanan 2017, Sun et al. 2016). These more advanced species-delineating methods have updated accepted phylogenies, which generally recognize four clades in the genus *Ulva*: *ohnoi/rigida*, *pseudocurvata/compressa*, *intestinalis/lactuca* and *linza/prolifera* (Figure 1.10). This work has set

the stage for more focused research, first through *Ulva* genome sequencing and now through developing the *Ulva* “molecular toolkit”, which includes transgenic, gain-of-function *Ulva* lines (De Clerck et al. 2018, Blomme et al. 2021).

Such advances have also assisted in the correction of historical records. Measures of species diversity based on morphological identification have been revisited with this genetic approach, resulting in updates to species prevalence and establishment ranges. The most recent research in this field shows disagreement between morphological-based species definitions used historically and molecular marker-based identifications (Tran et al. 2022). In Ireland, southern Britain, and the northwest Atlantic, historical surveys based on morphological and cytological data underestimated the species diversity by up to 66% and identified lineages which were not acknowledged in previous comprehensive evaluations (Burrows 1991, Loughnane et al. 2008, Hofmann et al. 2010).

1.3.2 *Ulva* life cycle

In the past decade, groups have utilized the early development and growth of *Ulva* to create sustainable laboratory stock cultures to facilitate research. *Ulva* reproduces via alternation of generations, in which subsequent generations arise from mature gametophytes or sporophytes (Figure 1.11).

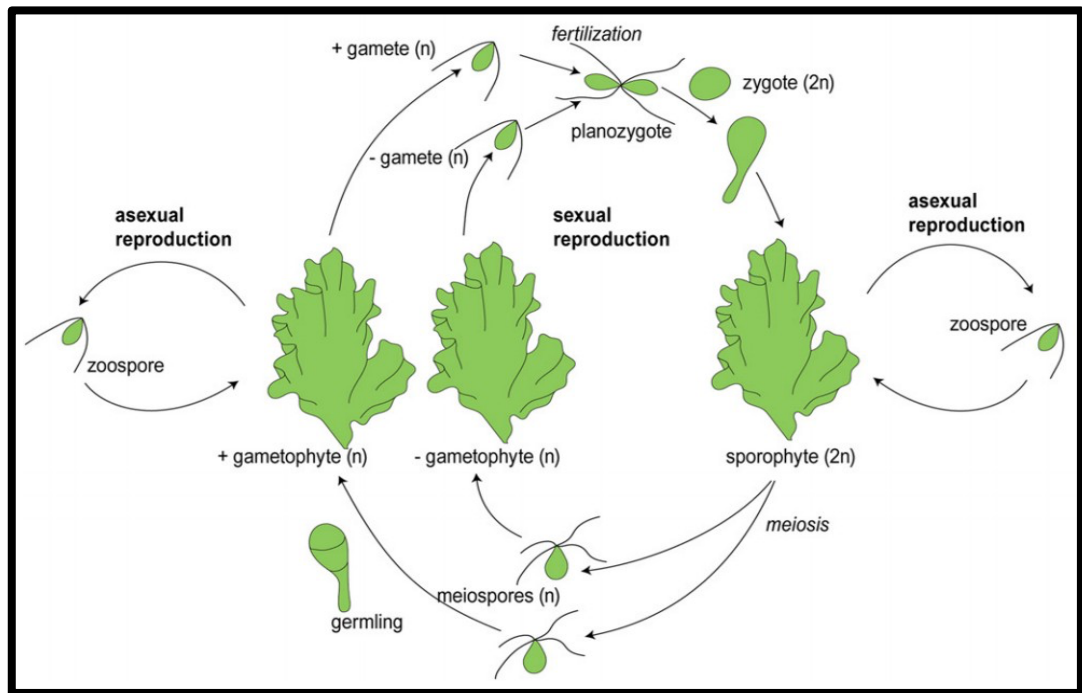


Figure 1.11 Life cycle of *Ulva* spp. Generations alternate between the haploid gametophyte and the diploid sporophytic forms (Leliaert et al. 2012).

Diploid sporophytes develop from paired biflagellate gametes while haploid gametophytes develop from quadriflagellate zoospores (Figure 1.11). Both forms can also arise parthenogenetically, and the mature frond phenotypes are virtually indistinguishable (Phillips 1990, Hoxmark 1975, Brodie et al. 2007). The presence of an extracellular sporulation inhibitor was discovered to maintain a vegetative state in the frond, preventing the onset of this gametogenesis (Stratmann et al. 1996). Removal of this inhibitor by fragmentation of the thallus allows for collection of gametes, if the thallus is a mature gametophyte, or zooids, if the thallus is a sporophyte (Wichard and Oertel 2010). The synchronous collection of the subsequent generation is ideal for developmental experiments and has been utilized recently in the discovery of the role of bacteria in early development through excretion of regulatory factors (Wichard and Oertel 2010, Spoerner et al. 2012, Vesty et al. 2015).

1.4 Seaweed genomics

Seaweed genome studies lag behind those of other more well established model organisms due in part to the nature of the seaweeds themselves. Their genome size is relatively small compared to the most widely used model organisms (*Drosophila* 180 Mbp, *Mus* 2,500 Mbp), which is an advantage. There are, however, unique challenges.

Notably, in *Ulva* diploid sporophytes and haploid gametophytes are visually indistinguishable. This is an obvious complication to any sequencing effort. Additionally, the ecological, morphological, and cytological diversity of seaweeds are reflected in their genomes. Many fully sequenced species have upwards of 35% functionally uncharacterized genes (Nakamura et al. 2013, Lin and Qin 2014). Genome sizes and gene contents also vary greatly between species. For example, *Saccharina japonica*, a commercially important brown algae, contains over 35,000 genes, while *Ulva mutabilis* contains only around 10,000 (Ye et al. 2015, De Clerck et al. 2018).

Some seaweeds, including *U. intestinalis* and *U. mutabilis*, interact obligatorily with resident marine bacteria (Wichard et al. 2015, Ghaderiardakani et al. 2017). Thus, laboratory specimens may not be successful when maintained in axenic cultures and field collected specimens must be treated before sequencing, with data being edited post-sequencing to remove contaminating bacterial sequences.

Lastly, *de novo* genome sequencing of these organisms often stops at the resolution of contigs and scaffolds. Contigs are formed by the overlay of short sequence fragments, but nucleotide polymorphisms, un-sequenced genomic regions, and repetitive sequences can affect the final alignment (Nagarajan and Pop 2013, Shi et al. 2016, Du et al. 2017, Jiao et al. 2017). Genome scaffolds refer to arrangements of contigs, which may be constructed by incorporating the distance information contained in paired end

reads. The mapping of scaffolds onto chromosomes is often not undertaken for genome assemblies generated rapidly from Next Generation Sequencing (NGS), because of considerations of time, cost, effort, and payoff.

Given that this is the case, most algal genomes exist as draft assemblies, where gaps exist between uninterrupted expanses of DNA sequence. Although genome completeness is often near 100%, these gaps are what limit these genomes from dropping “draft” as a caveat.

1.4.1. *Ulva* genomes

Sequencing of alternative plant model species is a fast-developing field. In the last three years alone, three notable additions to green algal genomics were sequenced and published: *Ulva mutabilis*, *Ulva prolifera*, and *Caulerpa lentillifera*.

The first of these was *U. mutabilis*, published as a reference genome in 2018, which represents a non-seeding plant model (De Clerck et al. 2018). This work focused on dissecting the *U. mutabilis* genome from an evolutionary perspective, specifically within the scope of multicellularity and the gene family expansions necessary to accommodate such a transition. Although this reference genome has since been published, the work completed in this thesis between 2016-2018 laid groundwork for which the direction of the genome publication eventually followed, an effort that was concerted in large part by supervisor, Dr John H. Bothwell.

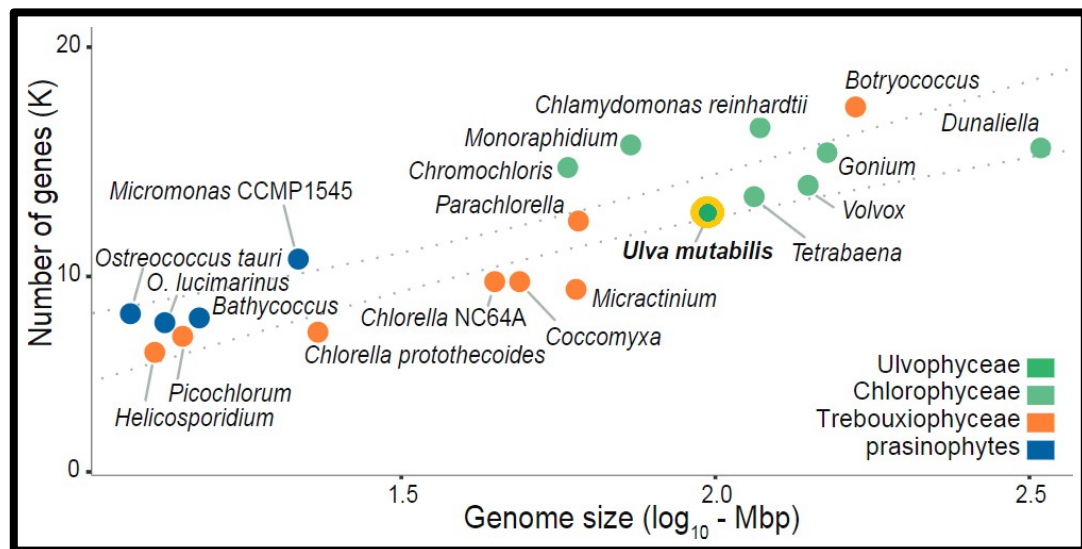


Figure 1.12 Algal genome comparison of relative genome size and number of genes between three core chlorophyte groups (Ulvophyceae, Chlorophyceae, and Trebouxiophyceae) and unicellular, flagellated prasinophytes. *Ulva mutabilis* is highlighted in yellow (adapted from De Clerck et al. 2018).

Compared to its close relatives, *U. mutabilis* has a mid-range sized genome and associated gene number (Figure 1.12). The published genome is presumed to be a largely complete dataset with predicted genome integrity of 98.5% and an estimated minimum of 92% genome completeness based on the eukaryotic BUSCO dataset. Of the nearly 13,000 predicted genes, 91.8% were supported with RNA sequencing data. As such, *U. mutabilis* is a prime subject for comparative analysis among emerging chlorophyte genomes, representing a reliable and largely complete close relative for which to compare the newly created assemblies proposed in this thesis (Chapter 2) as well as other recently sequenced green algal relatives. It is important to note that studies published prior to 2018 may refer to *U. mutabilis* and *U. compressa* as separate species. It was this year that a study published in the Journal of Phycology presents these species as conspecifics based on morphological patterns of vegetative and reproductive features, gamete compatibility, and molecular phylogenetics (Steinhagen et al. 2018).

More recently, Yuan He and associates from Soochow University published the *Ulva prolifera* genome (2020). This data represents the second sequenced *Ulva* species alongside the *Ulva mutabilis*/ *Ulva compressa* draft reference genome; an important addition to the publicly available knowledgebase on which to further develop our understanding of green algal biology via comparative analysis.

After DNA extraction from field-collected samples, the Soochow group automated *de novo* assembly and annotation using a variety of pre-packaged software programs (Chin et al. 2016, Schattner et al. 2005, Lagesen et al. 2007, Stanke et al. 2004, Buchfink et al. 2015, Wickstead et al. 2003). These pipelines produced a 87.9 Mb genome with approximately 10,000 genes and a BUSCO integrity score of 75.3% (Simão et al. 2015). These data indicate that this study, while worthwhile, is lacking approximately 25% of the *Ulva prolifera* genetic code, a considerable proportion. Nonetheless 75% is significant to interrogate interspecific genome patterns within the *Ulva* genus.

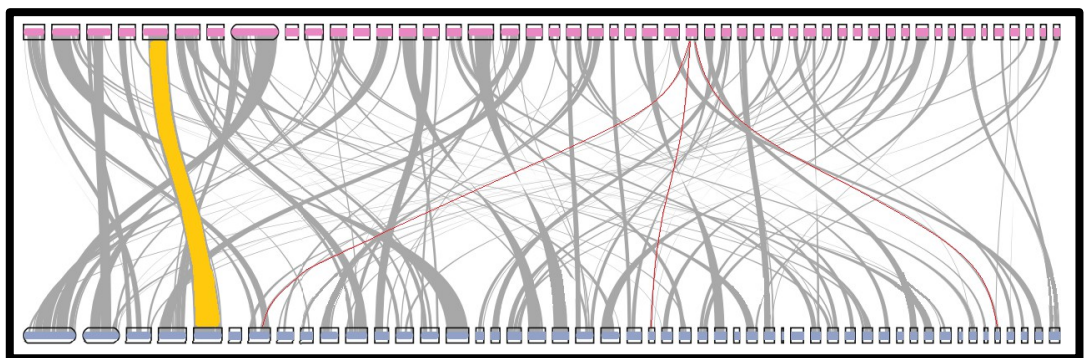


Figure 1.13 Collinearity analysis chord diagram with *U. prolifera* (pink) and *U. mutabilis* (blue), showing areas of synteny (yellow) and more diverse genomic regions (red)(adapted from He et al. 2020).

He al. touched upon this point, using collinear analysis (Figure 1.13) and a phylogenetic tree based on the whole genome of twenty green algal species. Their analysis shows agreement with many proposed phylogenies, mainly the widely

accepted notion that members of Chlorophyceae and Ulvophyceae are close relatives within green algal GenBank species. Notably lacking in this simple presentation of the *U. prolifera* genome are more concrete comparisons at the gene level.

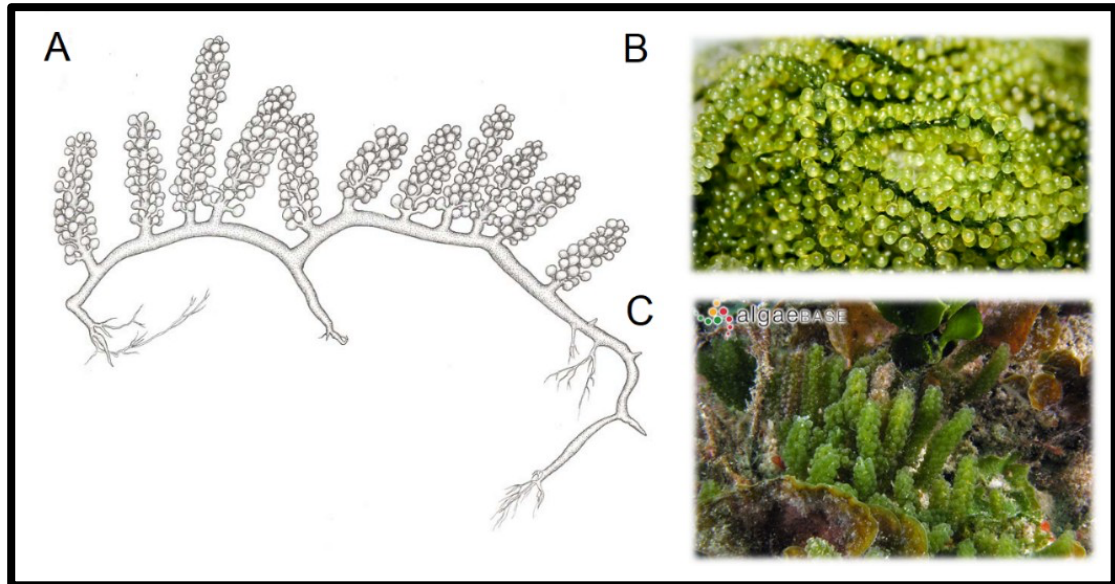


Figure 1.14 *Caulerpa lentillifera*, commonly referred to as ‘sea grapes’; (A) diagram of body plan, (B) close up of bulbous, grape-like structures, and (C) in its natural habitat, the intertidal (accessed via Algaebase 2021).

Another species of ulvophyte seaweeds, *Caulerpa lentillifera*, was sequenced by Arimoto et al. in 2019. Unlike *Ulva*, the genus *Caulerpa* represents a group of unicellular, often multinucleate, and siphonous macroalgae (Figure 1.14). This represents the third and final green macroalgal genome currently available. A time-calibrated molecular phylogeny of *tufA* and *rbcL* DNA sequences produced in 2014 determined an estimated distance of approximately 600-800 million years between *U. mutabilis* and *C. lentillifera* (Draisma et al. 2014). The higher end of this range (800 my) is supported by more recent studies of amino acid alignment of nuclear genes (Del Cortona et al. 2020).

Both morphologically and functionally, *Caulerpa* represents an evolutionarily distant divergent relative of *Ulva*. The work of Arimoto et al. (2019) suggests the Ulvophyceae, including *Ulva* sp. and *Caulerpa* sp., are not a monophyletic group (Figure 1.15). This is in contradiction to other recent studies that concluded the opposite (Sun et al. 2016, Melton et al. 2015). Obviously, this relationship warrants additional attention.

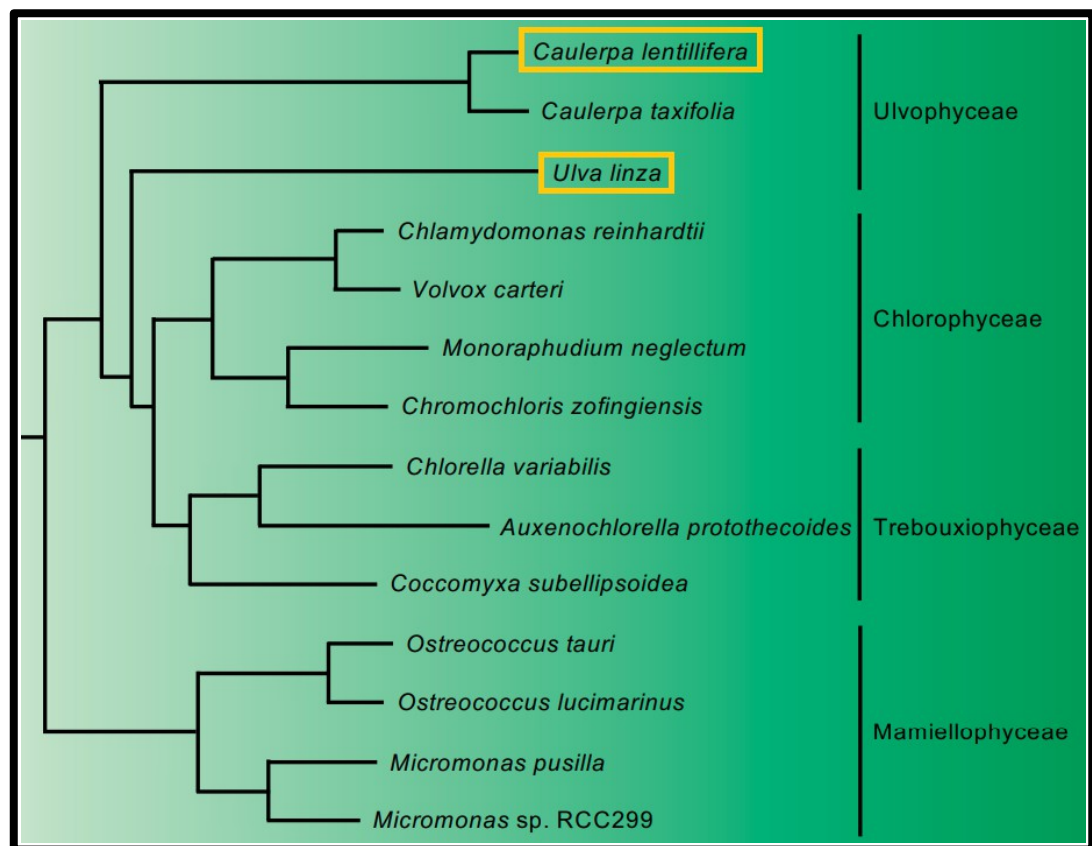


Figure 1.15 Phylogenetic tree of *Caulerpa* relationship to other chlorophytes. Tree was constructed with optimal maximum-likelihood method using a concatenation of nuclear-encoded protein alignments. *C. lentillifera* and *U. sp.* are highlighted in yellow (adapted from Arimoto et al. 2019).

Interestingly, the *Caulerpa* genome is smaller than the two *Ulva* genomes (28.7 Mbp) but contains a similar number of predicted protein-coding genes (9,311). These data are supported by a relatively strong BUSCO integrity score of 86.4% and are backed by RNA-seq data, so this is likely to be a real effect, although the explanation remains speculative.

1.4.2 Future directions

The flux of newly available genome data offers many opportunities for comparative analysis. Several recent studies focus on just that.

For example, Del Cortona et al. (2020) takes a geological record approach to understand diversification events in early macroalgae, with specific attention to multicellularity, macroscopic growth, and transition to benthic environments. By overlaying a time-calibrated phylogeny of green algae onto a geologic timeline and fossil data, Del Cortona et al. (2020) identify events marking evolutionary advancements in green macroalgae.

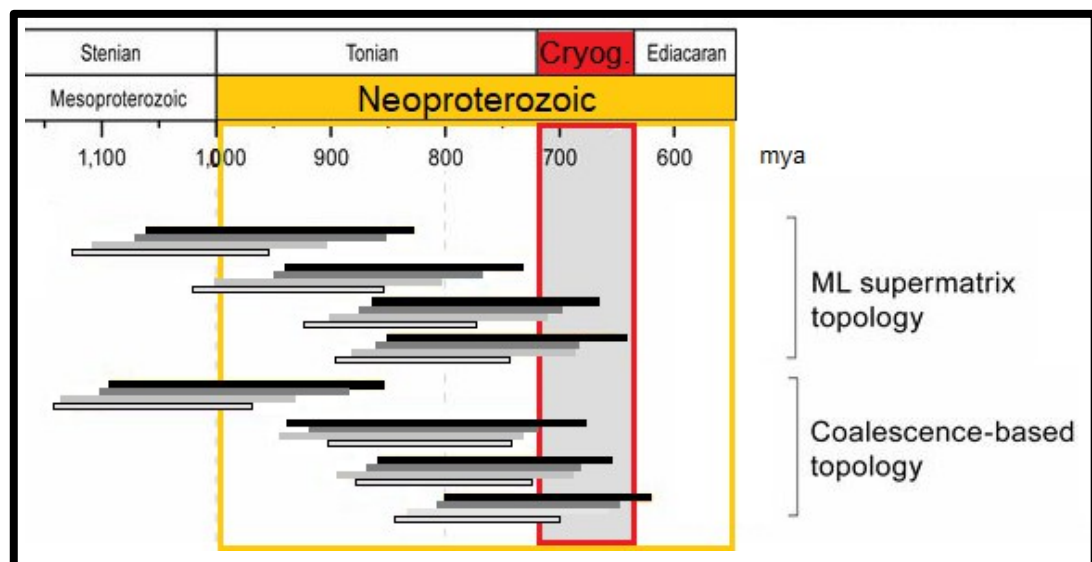


Figure 1.16 Divergence time calculated from time-calibrated phylogeny with two different methods (right). Grey bars represent estimates with different initial calibration assumptions. Yellow highlights the Neoproterozoic era and red marks the nested Cryogenian era (adapted from Del Cortona et al. 2020).

Of course, relying on fossil data is highly subjective to discriminative morphological features, which we already know to be a difficult parameter for delineating lineages within Chlorophyta, and *Ulva* in particular. However, these data accompany molecular clock-associated models, and indicate that the core Chlorophyta (Chlorophyceae, Bryophyceae, and Ulvophyceae) emerged 1,000 to 700 million years ago. Based on these data, the Ulvophycean lineage is believed to have radiated at the more recent end of this range, in the Cryogenian (Figure 1.16).

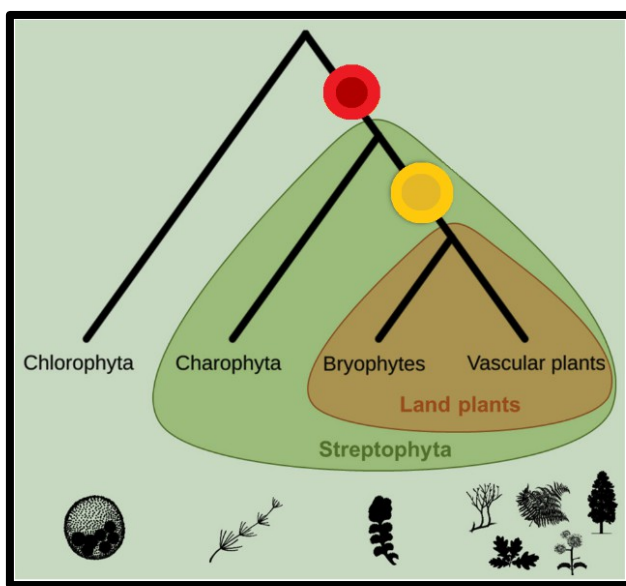


Figure 1.17 Simplified model indicating additions of “genomic novelty”, shown in red and yellow, hypothesized to be crucial in the evolution of land plants from the last common ancestor shared with the chlorophytes (leftmost taxa) (adapted from Bowles et al. 2020).

Interestingly, this diversification period overlaps with two major cooling glaciation events, suggesting an added selective pressure, and which is proposed to mark a time of return to benthic and coastal marine environments. This model fits with the proposed divergence of *Ulva* sp. and *Caulerpa* sp. by Draisma et al. (2014). Additionally, the resulting phylogenetic tree from the Del

Cortona et al. (2020) analysis supports the notion of Ulvophyceae as a non-monophyletic group, with statistical polytomy suggesting otherwise, as in Sun et al. (2016), Melton et al. (2015), and Arimoto et al. (2019). Lastly, but importantly, this phylogeny confirms the hypothesis that macroscopic growth evolved convergently in separate Ulvophycean lineages from an ancestral unicellular green alga; the mechanisms of which are predicted to involve multicellularity coupled with body

design (for example, siphonous) or cell division (for example, mitosis or cytokinesis) (Coquyt et al. 2010).

A second recent paper focused on green seaweed genomics characterizes novel genomic traits which predate the evolution of land plants. The ability of primary producers to leave water and colonize land is a dramatic functional leap. Bowles et al. (2020) hypothesizes the biological processes, and corresponding genomic novelties, required to accomplish such a transition.

Their deceptively simple model highlights two nodes of particular interest: one in the Ediacaran, involving new genes for multicellularity and leading to the rise of Streptophytes, and the other in the Ordovician, introducing new terrestrialization genes that prepared future land plants for a life outside of water. The timing of these events, particularly in the Ordovician, fit well with the so-called Great Ordovician Biodiversity Event, sometimes referred to as GOBE. As primary producers expanded to new habitats, they created novel physical complexity in their environments. Fauna co-evolved to fill the surplus of newly available ecological niches, particularly in marine environments (Servais and Harper 2018).

Bowles et al. (2020) used gene ontology from the well-annotated proteome of *Arabidopsis thaliana* to explore functions of extant descendants from core homology groups. They found that homology groups in the last common ancestor of land plants (yellow circle, Figure 1.17) were skewed towards protein modification. Meanwhile, homology groups present in the Streptophyte ancestral lineage (red circle, Figure 1.17) were rich in gene regulation, plant-specific transcription factors, root and host microbe interactions, lateral organ development, and a more complex cell wall structure compared to Chlorophytes.

Two possible mechanistic explanations were offered for this expansion of conserved genomic regions: whole genome duplication and/or horizontal gene transfer. The former was considered outside the scope of this paper, as a whole genome duplication event was not identified in the last common ancestor of the selected study groups. The latter was identified as a likely origin of expansion among gene regulation and protein modification homology groups within the parameters of this study.

Building on these results, investigations of genome change have also looked at variation in the *Ulva* chloroplast genome, both intra- and inter-specifically (Liu and Melton 2021). Differences in genome size in 13 species were attributed to intergenic spacers, foreign DNA integration, and, most notably, group I and II intron variation. These introns appeared to be dispersed more or less randomly, consistent with their previously defined role as mobile genetic elements (Bonen and Vogel 2001, Novikova and Belfort 2017). Gene identity was highly conserved with the same 100 genes present in each chloroplast, with the exception of one individual. What is interesting here is the support of the chloroplast genome in the amazing plasticity of these organisms. Although content is conserved, structure is not, due to rearrangement events from mobile elements. Chloroplast genome sequencing is more common compared to the whole genome counterpart and there is extensive data to compare among these sequenced individuals (Fang et al. 2017).

This flood of newly available information sheds new light on the field of whole genome *Ulva* comparisons. Specifically, the automated production and annotation of the *Ulva prolifera* genome lacks context and detail that may be better resolved with a more tailored, comparative approach. The BUSCO genome completeness score indicates 25% of data is missing from the genome, well below the standard benchmark of 5% for usability. However, as it was constructed by use the reference genome

U. mutabilis, a comparative analysis of the 75% present data is warranted, and provides a clear directive for this thesis

1.5 *Aims and objectives*

The central aim of this thesis is to investigate genomic variation in green macro algae of the genus *Ulva* with specific focus on species-level diversity via comparison of whole genome structure/content and to identify features that may underlie adaptation among these organisms.

The aims have been accomplished by:

- Generating novel genomic assemblies from field collected specimens.
- Comparing genomic structure and gene content across *Ulva* species.
- Observing patterns of gene orthology across Chlorophyte species, with particular attention to sulfotransferase-related genes.
- Identifying evolutionary patterns to explain diversity in sulfotransferase gene families.
- Establishing a foundation for future comparative studies of evolution, adaptation, and species relationships in *Ulva* and other Chlorophytes

Chapter 2: CREATION OF *DE NOVO* ASSEMBLIES

2.1 Introduction

Prior to the publishing of the *U. mutabilis* genome (De Clerck et al. 2018), the search continued for a simple and reproducible procedure to reliably extract high quality genomic DNA with little to no exogenous contamination. The idea was to create an accessible protocol for collaborators to use with local *Ulva* samples.

We deemed the simplest, most accessible, and most intuitive approach for DNA extraction was by way of field collected thalli. This method introduces inherent complications; most notably, prokaryotic contamination and inclusion of chloroplast DNA of morphologically identified species, which we know to be tedious and unreliable. Steps were added to what would be considered a typical DNA extraction protocol in specific consideration of these complications. Specimens were hand washed and gently scrubbed of visual epiphytes in an attempt to limit prokaryote DNA contamination. Tissue was washed with germanium, known to inhibit diatom growth, a food source of crustacean grazers, should they remain after washing (Thomas and Dodson 1974, Costa et al. 2016). Before extraction, thallus was visually inspected via microscope to confirm the sample was free of extraneous organisms and debris. A density-dependent centrifugation step with caesium chloride was used to stratify plastid and chloroplast DNA and purify the target genomic DNA (Thiery et al. 1976, Cooney and Matthews 1984, Richards et al. 1994).

The usability of this data is a primary focus in addressing questions of whole genome sequencing of field collected samples. We aim to determine whether our considerations of the unique nature of *Ulva* field specimens is sufficient to produce quality genomic DNA, free of microbial DNA, and of suitable completeness and quality to perform inter-specific comparisons.

The basic genomic comparison of regionally co-occurring *Ulva* species is yet unexplored. In this chapter I developed two genomic assemblies from the *Ulva* genus, partially covering the entire genome and henceforth referred to as pseudoassemblies, which were created with the intention of comparison to the newly published (at the time, 2018) reference genome. Here we reflect on the efficacy of this methodology, the quality of the resulting data, and develop a dot plot-based comparative methodology for analyzing related genomes.

2.2 Methods

2.2.1 *Sample location and identification*

Two locations, Castle Ward Shoreline in Northern Ireland and Langstone Harbour in Southern England (henceforth referred to as Ward and Langston, respectively), were selected based on their relative geographic isolation and native *Ulva* populations. Several *Ulva* species are well documented throughout the coastline of the British Isles, including these distinct regions (Figure 2.1).

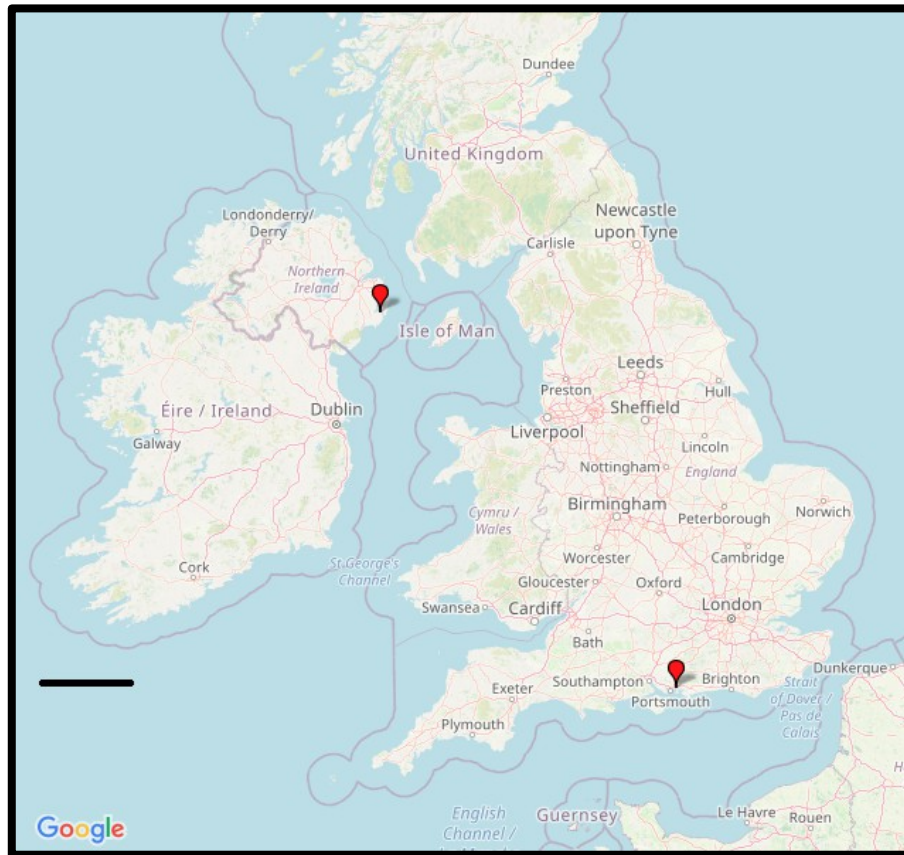


Figure 2.1 Sample collection from two locations sites in the United Kingdom: Ward (upper red point) and Langston (lower red point). Scale in the lower left equal to 100km (Google).

Several samples of attached thalli were collected from each location and transported for culture in the laboratory. Artificial seawater (ASW) media was prepared for laboratory culture of the unique, non-axenic field specimens. ASW was fortified with germanium to eliminate epiphytic diatoms, a technique known to disrupt diatom cell wall structure by replacing biologically available silica (Markham and Hagmeier 1982).

The most robust and morphologically similar samples from each culture were identified and isolated. A central square of thallus with mass greater than 50 mg and less than 100 mg was removed, which equates to an approximate area of 1 cm² (Figure 2.2). The remaining thallus was desiccated and stored.

2.2.2 Extraction and purification

Thalli samples were homogenized with mini-pestles and placed on ice. DNA was extracted according to protocol set forth in DNeasy plant mini kit (Qiagen) with specific parameter changes based on a series of trial extractions (Chapter 6).

Following elution, equilibrium density gradient centrifugation was performed with caesium chloride to isolate *Ulva* sp. DNA from bacterial-symbiont DNA. Ultimately, this technique allows for density-specific separation of circular bacterial DNA and sheared, linear eukaryotic DNA, a critical step

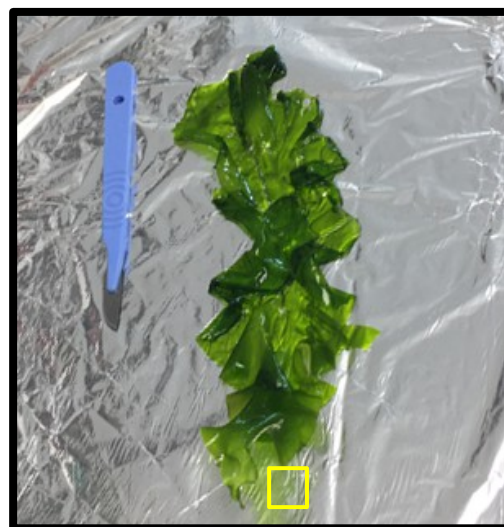


Figure 2.2 Culture sample in preparation for DNA extraction. Hypothetical extraction sample highlighted by yellow square.

considering the facilitating role of marine bacteria in normal *Ulva* morphogenesis (Carr and Griffiths 1987, Spoerner et al. 2012).

2.2.3 Sequencing

Illumina single-end, 150 base pair, whole genome sequencing was performed. Two replicates were run for each the Ward and Langston samples. The resulting dataset represents samples of four individual algal specimens identified morphologically as members of the genus *Ulva* from two geographically distinct regions.

2.2.4 Quality Analysis

From the set of data files obtained from the shotgun sequencing runs, four raw fastq files were identified as the input datasets, which represented two individuals isolated from each Ward and Langston. Linux served as the operating system for the quality analysis and trimming of the four raw fastq files. Quality assessments and trimming

were performed on each individual file using the Trimmomatic function with the following parameters: sliding window 4:15, minimum length 36 base pairs, single end reads (Bolger et al. 2014). Over a span of 15 consecutive bases, each must have a sequencing depth of four or greater, otherwise the sequence was omitted. This process eliminated sequences that were left inadmissibly short, i.e., less than 36 bases in length, or of insufficient quality.

2.2.5 *De novo Assembly*

Using the Linux operating system, the genome was indexed using BMap (Bushnell 2014). The SSake function was used to create *de novo* assemblies from the quality-assessed files (Warren et al. 2007). The Flagstat function of SAMtools bundle reported summary statistics of the assemblies including GC enrichment, N/L50, N/L90, number of contigs, contig length, total sequence length, and contig size bins (Cock et al. 2015).

2.2.6 *Alignment*

The BWA tools short read alignment function was then used to align the quality assessed fastq files to the reference genome. The reference genome for this study was *Ulva mutabilis*, the first sequenced and most comprehensively understood genome from the *Ulva* genus (De Clerck et al. 2018). The flagstat function provided summary statistics detailing these alignments, which required conversion of the output to BAM files.

The results of the initial alignment were exported using the Mpileup function in the SAMtools toolkit. The summary output was produced in the form of ‘.txt’ files, which were converted to a searchable excel database containing exact nucleotide matches to the *U. mutabilis* draft genome, along with a quality score, and the location of the match in the reference genome.

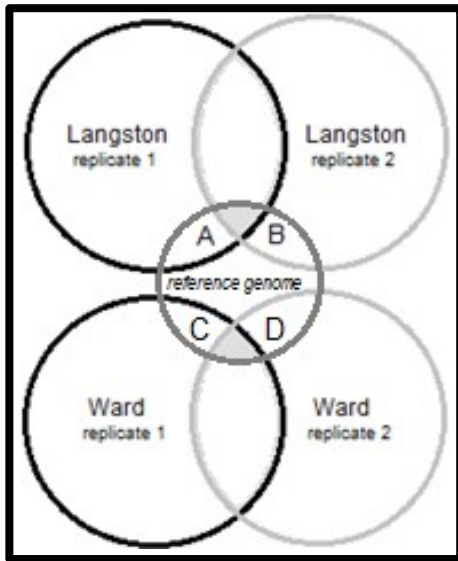


Figure 2.3 Model depicting SNP identification resulting from the Mpileup analysis. SNPs are the result of a match between one replicate (either A and B in the case of Langston, or C and D in the case of Ward) and the reference genome, highlighted in grey subsection.

As a measure of diversity between each replicate used to create *de novo* pseudoassemblies, single nucleotide polymorphisms (SNPs) were tallied between replicates in both Ward and Langston from the Mpileup extraction (Figure 2.3). These SNPs represent the subset of bases with likeness to the reference genome in one replicate of each experimental sample. This excludes three-way matches between each replicate and the reference genome, *Ulva mutabilis*.

A second subset of this database was created by filtering the original data to show only nucleotide bases which were present and consistent between both replicates from each location, in addition to the reference genome (Figure 2.4). This is a three-way comparison encompassing agreement from each replicate of each location and the reference genome. Single nucleotide matches from the two replicate datasets for each location, Ward and Langston, and the reference genome were concatenated to form continuous sequences from the list of matches.

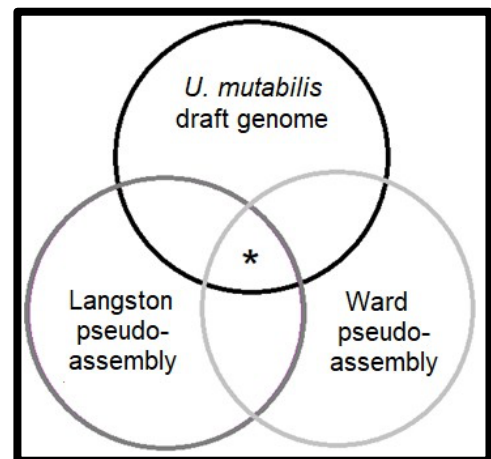


Figure 2.4 Model depicting resulting sequencing extraction from the Mpileup analysis, highlighted by central asterisk.

This method functionally eliminates SNPs from the previous analysis and other unique sequences but allows for a comprehensive view of the similarity between individual one and two from each study location and the reference genome. Additionally, perfect one, two, and three-base repeat sequences were eliminated from the resulting data to favor more unique and conserved genomic regions. For the resulting list of contiguous nucleotide sequences, the size in base length, location within the reference genome including starting position, and average count for each experimental sample were identified (Digital Appendix 1.1-1.2).

2.2.8 Outgroup annotation

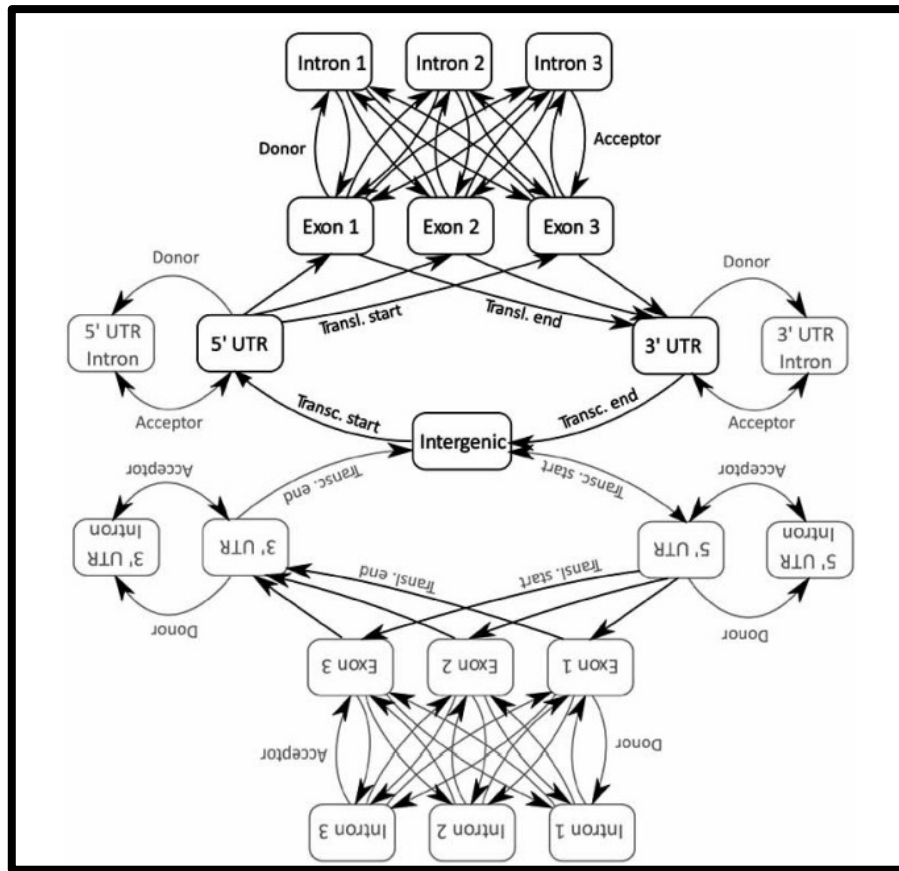


Figure 2.5 Possible annotation detection pathways via the EuGene platform (Foissac et al. 2008).

The annotation software *Eugene* was used as a comparison tool between Langston or Ward and a distant relative, *Arabidopsis thaliana*, to verify the presence of conserved sequences not only in a close relative, as in *U. mutabilis* (Foissac et al. 2008). Contigs greater than 1,000 bases in both Ward and Langston pseudoassemblies were targeted, of which there were a total of 26; 11 from Ward and 15 from Langston (Table 2.1). Predictive genes and gene features were annotated from the reference of the *A. thaliana* genome on EuGene using the detection pathways presented (Figure 2.5).

2.3 Results

2.3.1 Assembly construction

Both pseudoassemblies were constructed using the Linux operating system, which was accessed via the Durham University Hamilton server, using a series of operations designed for quality control and *de novo* assembly from four respective input fastq files. The entire sequence read workflow can be tracked from beginning to end for Ward and Langston, respectively (Figures 2.6 and 2.7).

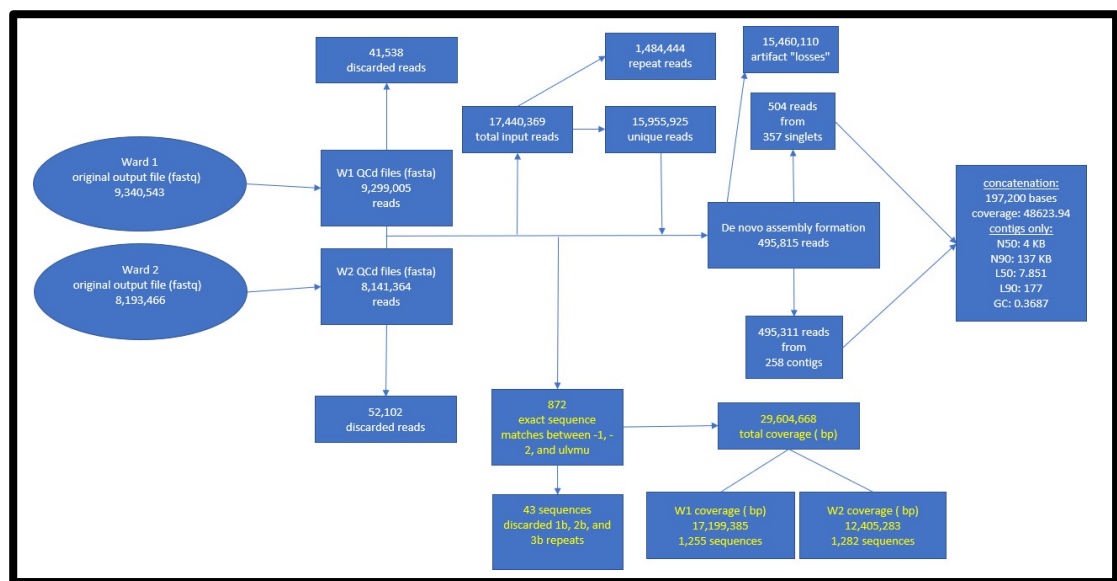


Figure 2.6 Read workflow for Ward files, from original fastq to *de novo* pseudoassembly, with counts and demographics. Yellow text indicates demographics from a second, searchable database comparing the reference genome with each individual replicate.

With respect to the Ward dataset, approximately 0.5% of the 17.5 million input fastq reads were discarded following Trimmomatic with the parameters previously described. Of the remaining 17.4 million reads, 15.9 million represented unique sequences, with 1.5 million accounting for reads of repeating sequences. The white text in figure 2.6 represents the following workflow. The final assembly was created with 495,311 reads, 2.8% of the original quality-verified 17.4 million reads. These reads create 258 contigs and 357 singlets.

Artifact losses at 15.5 million indicates reads that contribute to depth, instead of unique

assembly sequence. Joining the contigs and singlets into a full pseudoassembly file covers 197,200 bases with a depth coverage of 48,624.

A separate alignment was created using Mpileup to determine identity to the reference genome, *U. mutabilis*, and SNP count between replicates for those sequences which match in the subsequent database (Figure 2.6). These data are represented in yellow text in figure 2.6. Of the entire set of replicate and reference data (9.3 million reads from Ward replicate 1, 8.2 million reads from Ward replicate 2, and the *U. mutabilis* reference genome), a total of 842 sequences were a 1:1, 100% match between the three. Of these matches, 5% were perfect one, two, or three base repeats, which were discarded to select for unique and more informative sequences. Depth of sequencing was tracked for each replicate individually (17.2 million for 1,255 sequences from Ward replicate 1 and 12.4 million for 1,282 sequences from Ward replicate 2) for a total coverage of 29.6 million bases. The complete Ward fasta file is available (Digital Appendix 2.1).

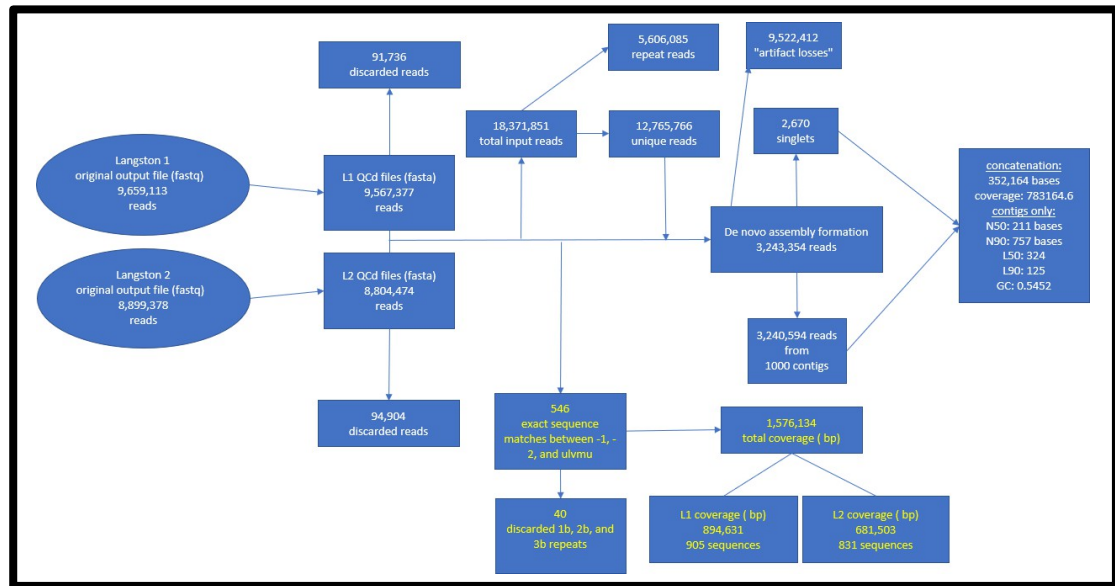


Figure 2.7 Read workflow for Langston files, from original fastq to de novo pseudoassembly, with counts and demographics. Yellow text indicates demographics from a second, searchable database comparing the reference genome with each individual replicate.

With respect to the Langston data set, approximately 1% of the 18.6 million input fastq reads were discarded following Trimmomatic with the parameters previously described. Of the remaining 18.4 million reads, 12.8 million represented unique sequences, with 5.6 million accounting for reads of repeating sequences. The white text in figure 2.7 represents the following workflow. These 18.4 million reads were reduced to 3.2 million reads, approximately 18% of the original quality-verified reads, in the *de novo* assembly. These reads create 1,000 contigs and 2,670 singlets. Artifact losses at 9.5 million indicates reads that contribute to depth, instead of unique assembly sequence. Joining the contigs and singlets into a full pseudoassembly file covers 352,164 bases with a depth coverage of 783,165.

A separate alignment was created using Mpileup to determine identity to the reference genome, *U. mutabilis*, and SNP count between replicates for those sequences which match in the subsequent database (Figure 2.7). Of the entire set of replicate and reference data (9.7 million reads from Langston replicate 1, 8.9 million reads from

Langston replicate 2, and the *U. mutabilis* reference genome), a total of 546 sequences were a 1:1, 100% match between the three. Of these matches, 7% were perfect one, two, or three base repeats, which were discarded to select for unique and more informative sequences. Depth of sequencing was tracked for each replicate individually (894,631 for 905 sequences from Langston replicate 1 and 681,503 for 831 sequences from Langston replicate 2) for a total coverage of 1.6 million bases. The complete Langston fasta file is available (Digital Appendix 2.2).

2.3.2 *Demographics*

Following the actual creation of the two *de novo* pseudoassemblies, summary statistics were recalled from the data via flagstat. The summary statistics serve as a tool to broadly understand the structure and composition of each pseudoassembly. The pseudoassemblies were compared to the reference genome *U. mutabilis* with the intention of evaluating their usefulness as novel data in the field (Table 2.1

Table 2.1 Basic statistics of de novo pseudoassemblies, Ward and Langston.

	Ward	Langston
Initial reads (replicate 1/ replicate 2)	17,534,009 (9,340,543 / 8,193,466)	18,558,491 (9,659,113 / 8,899,378)
Unique reads	15,955,925	12,765,766
SNPs (Mpileup)	73,968	77,186
Assembled reads (% of total)	495,815 (3.1)	3,243,354 (25)
Contigs (>250 bases)/ number of singlets	258 (69) / 357	1,000 (302) / 1,072
Number of bases/ coverage	197,200 / 48,624	352,164 / 783,164
N50 (bases)	4,000	211
L50	7.851	324
GC	0.3687	0.5452

The Ward pseudoassembly comprises of 495,815 reads, which is 3.1% of the total reads from the combined original fastq replicates. Of Ward's 258 contigs, 69 (28%) are greater than 250 bases. Ward is also comprised of 357 singlets ranging in size from 38 to 96 bases. These singlets represent sequences which have met quality control parameters, but do not overlap sufficiently to be considered a contig. These singlets range in coverage from one to three. This pseudoassembly includes 197,200 bases with a depth of coverage of 48,624. The conventional measures of N50 and L50 are 4,000 and 7.9. GC content is markedly lower in Ward at 37%, which is consistent with the low coverage of this genome.

Table 2.2 Contigs from Ward and Langston pseudoassemblies with base length 1,000 or greater.

Ward		Langston	
Contig	Length (bases)	Contig	Length (bases)
8	1193	37	1016
19	43549	44	1118
22	1443	50	3882
42	3238	84	1018
59	19593	135	2209
69	5718	139	1036
100	13831	192	1369
136	7318	304	1869
161	7852	331	1666
162	1459	418	1143
208	3444	497	1047
-	-	619	3977
-	-	923	1559
-	-	972	1092
-	-	996	1857

The Langston pseudoassembly comprises of 3,243,354 reads, which is 25% of the total reads from the combined original fastq replicates. Of the 1,000 contigs, 302 (30%) are greater than 250 bases. Langston also includes 1,072 singlets ranging in size from 40 to 96 bases. The upper end of this size range represents the median length of input sequences into the assembly workflow. These singlets represent sequences which have met quality control parameters, but do not overlap sufficiently to be considered a contig. The majority of singlets range in coverage from one to five, with several outliers between 10 and 14, and very few between 36 and 439. This pseudoassembly includes 352,164 bases with a depth of coverage of 783,164. The N50 and L50 are markedly lower compared to Ward, at 211 and 324 bases, respectively. GC content is markedly higher at 55%.

From the Mpileup, which again, represents exact nucleotide matches to the reference genome by position in at least one replicate file, Ward was determined to have 73,968 SNPs while Langston reported 77,186 (approximately 33 and 50%, respectively).

Bandage plots (Bioinformatics Application for Navigating De novo Assembly Graphs Easily) offer an alternative method of visualizing these data (Wick et al. 2015). The graphical depictions of each pseudoassembly show the count and scaled size of each contig in the data file. With these plots, we can better understand the number and size of contigs in each pseudoassemblies (Figure 2.8).

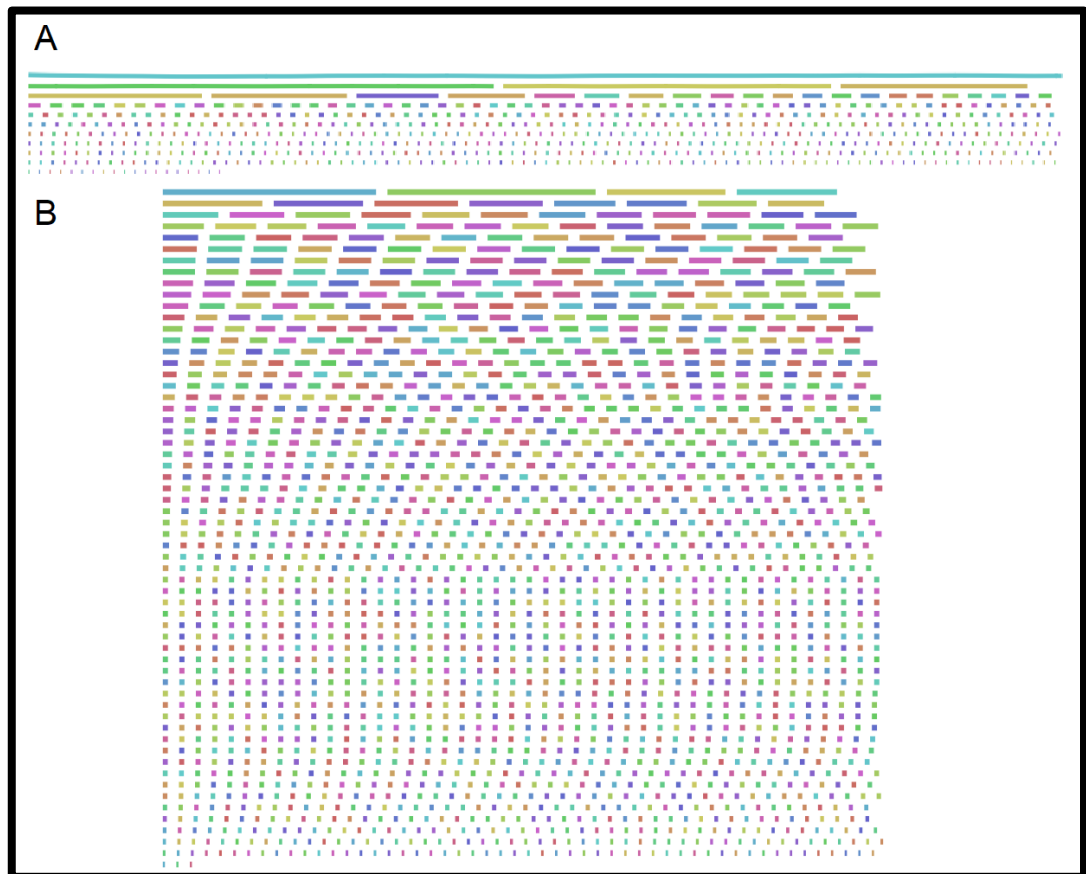


Figure 2.8 Bandage plots for (A) Ward and (B) Langston. The upper- and left-most turquoise band represents the length in bases of the longest contig in the assembly; (A) 43,548 and (B) 3,977 bases, respectfully.

The Bandage plot visual emphasizes the fundamental differences in composition between the two pseudoassemblies. The structure of Ward is heavily skewed by its

largest contig, which is 43,549 bases in length (Figure 2.8A). Aside from this relatively massive sequence, the remaining contigs follow a more predictable and familiar decreasing size pattern. Langston's largest contig is over an order of magnitude smaller at 3,977 bases (Figure 2.8B). Several relatively large contigs are followed by many smaller ones. Contig frequency distributions are available in Figure 2.10. While the demographic statistics of assemblies indicate overall poor coverage and quality, the question remains regarding their composition.

2.3.3 Assembly analysis

Dot plot analysis was performed via *Dgenies*. Each pseudoassembly was plotted against itself as a control measure, to ensure each pseudoassembly file was correctly formatted and consistent when compared against itself (Figure 2.9). Not pictured here, but also performed, was the reference genome control dot plot. Each dot plot was constructed by ordering contigs in ascending order on both the x- and y-axes.

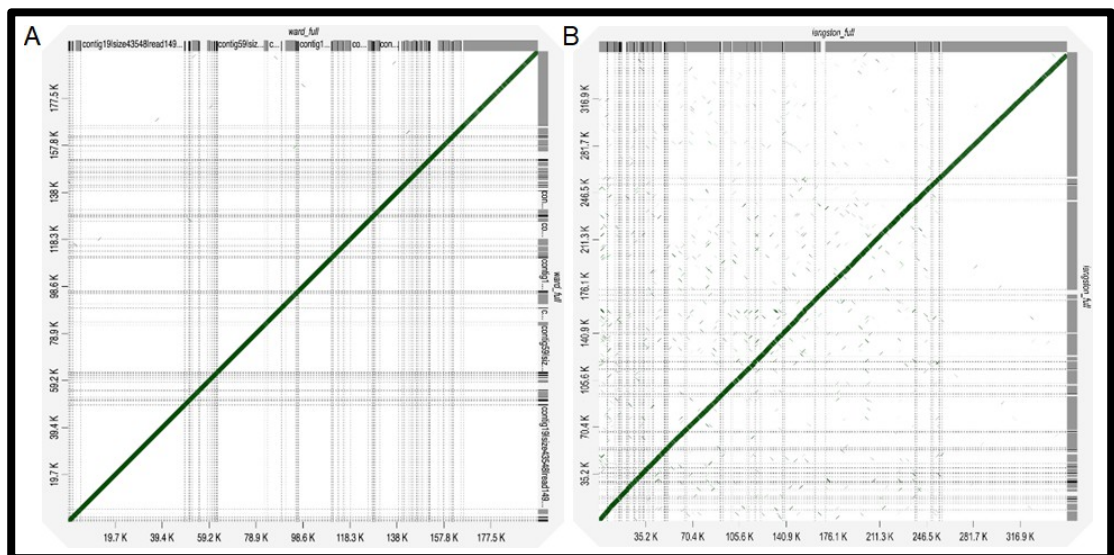


Figure 2.9 Control analysis dot plots of (A) Ward and (B) Langston pseudoassemblies plotted against themselves. Contigs are ordered ascendingly on both the horizontal and vertical axes.

The defined central diagonal in both the Ward graph (A) and Langston graph (B) indicates positional agreement from each axis. Any hits off the center diagonal are indications of repetitive sequences in the file, which is a normal artifact of these comparisons. Langston shows a higher number of these repeating sequences, indicated by the off-center, yet strong identifying, graphical marks.

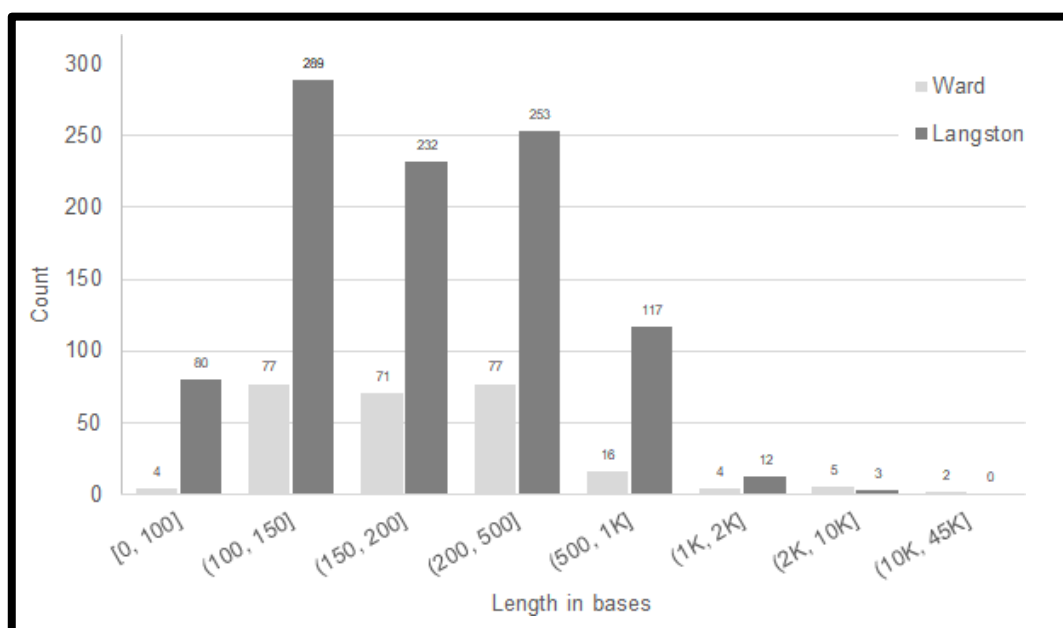


Figure 2.10 Contig length frequency distribution. Bins range in size from 100 to 45,000 bases. Ward contigs appear in light grey, Langston in dark grey. The count appears above the bar in each bin.

The annotation of contigs greater than 1,000 bases in each pseudoassembly (Table 2.2) were cross-referenced to the *A. thaliana* genome and resulted in several graphical and numerical outputs, which represent predictions of the query in the form of untranslated and intergenic regions, introns, exons, splice sites, and translation start and stop sites. These predictions were overlaid with a normalized likelihood Markov model, protein similarities between the target and query with gaps, and the prediction itself.

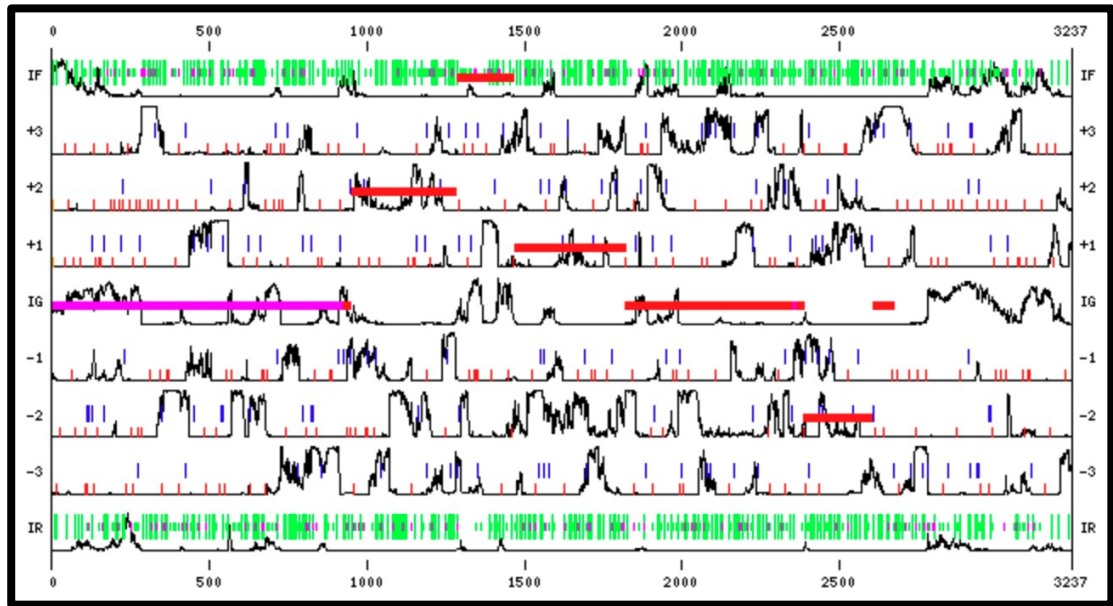


Figure 2.11 Ward contig 42 annotated alongside *A. thaliana* using Eugene gene finder software. The horizontal axis is the query sequence, and the vertical tracks are forward introns, coding +1-3, intergenic regions, coding -1-3 and the reverse introns, from top to bottom. The vertical blue and red lines are in-frame translation start and stops. The horizontal red and purple lines are the sequence predictions, representing intergenic and untranslated regions, respectively. The black plot lines are predictive Markov models.

In the Ward pseudoassembly, 45% of >1,000 base contigs produced predicted translation start and stop sites and Markov model, but did not identify with *A. thaliana* genes, nor predict any of those regions on the query sequence. Of the 55% which provided hits to the *A. thaliana* genome, the software was able to annotate both 3' and 5' untranslated regions and single exon genes in 100% of the proposed contigs. One Ward contig, contig 42, was also annotated with initial and terminal exons for the proposed identifying genes in *A. thaliana* (Figure 2.11).

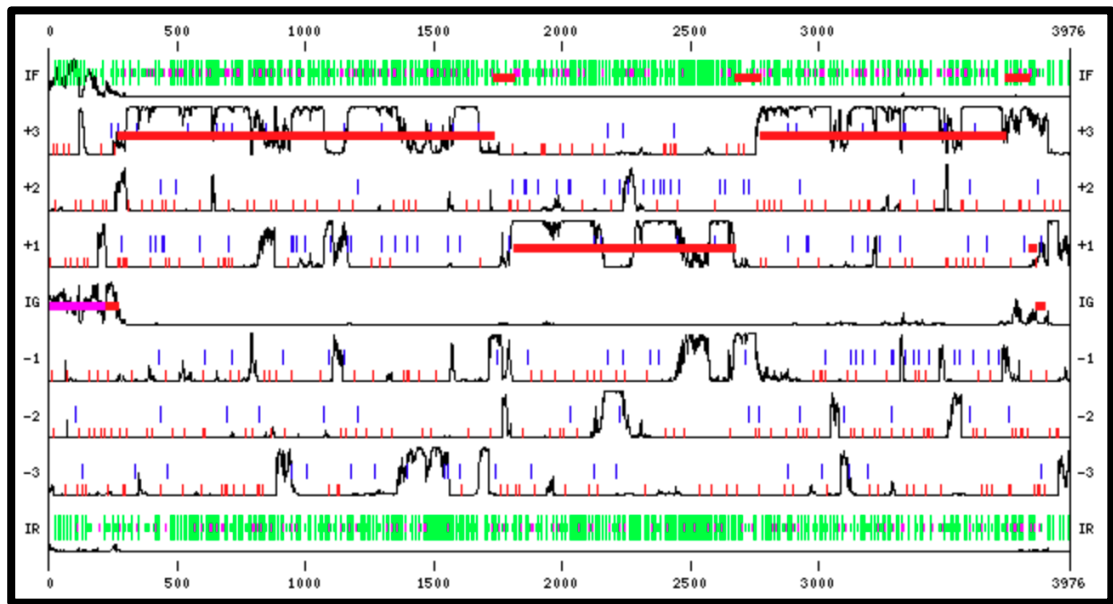


Figure 2.12 Langston contig 619 annotated alongside *A. thaliana* using Eugene gene finder software. The horizontal axis is the query sequence, and the vertical tracks are forward introns, coding +1-3, intergenic regions, coding -1-3 and the reverse introns, from top to bottom. The vertical blue and red lines are in-frame translation start and stops. The horizontal red and purple lines are the sequence predictions, representing intergenic and untranslated regions, respectively. The black plot lines are predictive Markov models.

In the Langston pseudoassembly, 20% of >1,000 base contigs produced predicted translation start and stop sites and Markov model, but did not identify with *A. thaliana* genes nor predict any of those regions on the query sequence. Of the 80% which provided hits to the *A. thaliana* genome, the software was able to annotate both 3' and 5' untranslated regions and single exon genes in 100% of the proposed contigs. Six Langston contigs were also annotated with initial and terminal exons for the proposed identifying genes in *A. thaliana* and one contig, contig 619, was also annotated with internal exons (Figure 2.12). This contig in the Langston pseudoassembly represents the longest of all contigs in the file, and presumably, contains the most usable data for comparative analyses.

Chapter 3: CHARACTERIZATION OF *DE NOVO* PSEUDOASSEMBLIES IN THE CONTEXT OF ESTABLISHED REFERENCE GENOMES

3.1 Introduction

The *U. mutabilis* reference genome was sequenced using gametes collected via induced gametogenesis (De Clerck et al. 2018). Part of the reason *U. mutabilis* has become a model for non-seeding chlorophytes is the relative ease of laboratory culture of this species and its readiness to differentiate upon cues conducive to artificial induction. The induction process involves fragmentation of mature thallus, followed by several washing steps intended to rinse tissue of sporulation inhibitors, which prevent differentiation in normal, healthy tissue. Genomic DNA can then be extracted from gametes, in relatively axenic culture media. This methodology has more recently been applied to transcriptomic research during different stages of development (Liu et al. 2021). Laboratory cultured *U. mutabilis* was the first specimen from which this behavior was characterized, and it has since been configured in a few other *Ulva* species (*linza* and *lactuca*) but does not appear to be a genus-wide feature (Wichard and Oertel 2010, Vesty et al. 2015).

Methodologies were designed in specific considerations of the harvesting, culture, and DNA extractions of *Ulva* thalli from which the Ward and Langston pseudoassemblies were created. The optimistic, yet expected, outcome was a relatively whole genome in which to use in comparative analysis. A consequential outcome was therefore evaluation of the efficacy of this particular methodology in genome DNA extraction of field-harvested live thalli.

Along with the recognition of the need for a non-seeding, multicellular plant model comes the opportunity to further develop the genetic toolkit of *U. mutabilis*, proposed after the thalli extraction work was underway (De Clerck et al. 2018). The distinctive

traits that make the *Ulva* genus a promising model system in phycology also contribute to the lack of clarity in how its species are defined and standardized. Most notably, *Ulva* are capable of a wide array of physical forms attributed to genomic plasticity and interspecific hybridization. Comparison of regionally diverse specimens is a primary focus in understanding the uniqueness of *Ulva*, and that is the ultimate fate of the Ward and Langston pseudoassemblies, which represent novel green algal genomic data of suitable quality, regardless of file size or relative genomic coverage. Here we present the outcome of this extraction method by comparing the resulting genomic content to the gamete-origin *U. mutabilis* reference genome.

3.2 Methods

3.2.1 *Comparison of relative structure and content*

Sequences with identity between each pseudoassembly and the reference genome, *Ulva mutabilis*, were extracted and annotated manually via ORCAE and BLAST (Sterck et al. 2012, Altschul et al. 1990). Assembly FASTA files were analyzed using dot plot via the platform Dgenies (Cabanettes and Klopp 2018). These visuals provide insight into repeating regions, breaks, inversions, and general genomic alignment. Pseudoassemblies were formatted to sort contigs into the best-fit alignment relative to the query reference sequence, as opposed to the arbitrary contig numbering system assigned during assembly creation. Matching sequences between both target and query were omitted from analysis if the sequence size fell between 0.001-0.2% of the dot plot width, which is scaled to genome size.

3.2.2 *Gene annotation*

High identity matches which fell outside of the diagonal identity match line on each dot plot were exported to a CSV file and manually extracted using the reference genome sequence in ORCAE (Sterck et al. 2012). The correlating assembly sequence of 80% match or better was analyzed via BLAST for similar sequences to verify if these regions are conserved among related species and genera and to identify gene ontology (Altschul et al. 1990).

Contigs of 1,000 bases or more from both Ward and Langston pseudoassemblies were compared to the reference genome *U. mutabilis* using BLAST. Perfect matches of 200 bases or more were cross referenced with the other pseudoassembly and annotated with gene function via BLASTx and ORCAE to determine the content of each sequence (Gish and States 1993). Discontiguous megablast was performed among highest identity reference genome contigs to determine contribution of each pseudoassembly to variation among a discrete nucleotide range.

3.3 Results

3.3.1 *Demographics and summary statistics*

Both pseudoassemblies represent a fraction of the scale of the reference genome, *U. mutabilis*, which is 98.5 Mbp in length from an original long read bank of 6.9 Gbp. Although it does not speak to the content of the assemblies or their usability as comparative tools, by length Ward covers 0.2% of the reference genome and Langston 0.4%. Ward is similar in number of contigs as compared to *U. mutabilis*; 258 and 318, respectively (Figure 3.1).

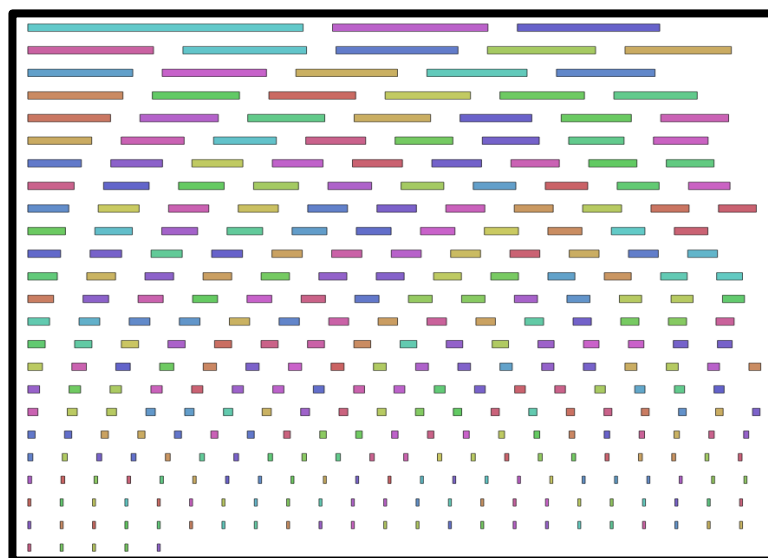


Figure 3.1 *Ulva mutabilis* bandage plot. The upper- and left-most turquoise band represents the length in bases of the longest contig in the genome; 3,623,364 bases. (Wick et al. 2015).

Langston has far more contigs as compared to *U. mutabilis*: 1,000. The proposed pseudoassemblies also contain singlets (Chapter 2). These sequences do not have a counterpart in the reference genome.

The visual representation of genome Ward compared to the reference genome shows a disparity in relative contig size and abundance. The collection of contigs in Ward (few large contigs and many small ones) is certainly a consequence of its largest contig, containing nearly 45,000 bases. The three next largest contigs are also in the range of thousands. Ward contains a similar number of contigs as compared to *U. mutabilis*, 258 and 318 respectively. But a N50 of 4,000, inclusion of singlets, and GC content of 36%, compared to 57% in *U. mutabilis*, does not support Ward as a strong comprehensive and comparative model.

Langston has a remarkably similar visual construction as compared to *U. mutabilis*, overlooking the vast differences in scale with Langston's largest contig at 3,977 bases and that of the reference genome at over 3.6 million bases (Figure 2.8). That is, the relative contig count and size distribution is similar. Langston also has a similar GC

content to the reference genome at 55%. Even so, with an N50 and L50 of 211 and 324, respectively, Langston is also not a strong candidate among whole genome assemblies compared to the reference genome's N50 of 527,412 bases and L50 of 54.

3.3.2 *Genomic content*

Assumptions of pseudoassembly quality based solely on comparative summary statistics is merited by the dot plot analyses comparing each pseudoassembly with the reference genome. Pseudoassembly Ward versus *U. mutabilis* presents an uncharacteristic dot plot for closely related species (Figure 3.2). Over 50% of reference genome contigs did not register likeness within Ward and were subsequently omitted from the figure presentation. Of those reference genome sequences which provided hits 50% or better on Ward, excellent coverage of several query sequences was observed: most obviously contigs 19, 26, 59, and 100. In a near opposite visual, few Ward sequences from single contigs appear to recur in many reference genome contigs; mainly Ward contigs 124 and 162.

Interestingly, and seemingly a strong confirmation of contig arrangement in the pseudoassemblies, it appears that contig order assignment by the assembly creation pipeline is not assigned completely arbitrarily. In many of the examples listed above, for example contigs 124 and 162 in Ward, hits to the reference genome span the delineation between flanking contigs. That is, hits along Ward contig 124, are in fact carried over continuously from the latter end of contig 123, and similarly for recurring hits in Ward 162, which are actually a continuous read spanning the end of Ward 161 through 162. This suggests quite definitely that the contig content, and in some cases order, are accurate and reliable.

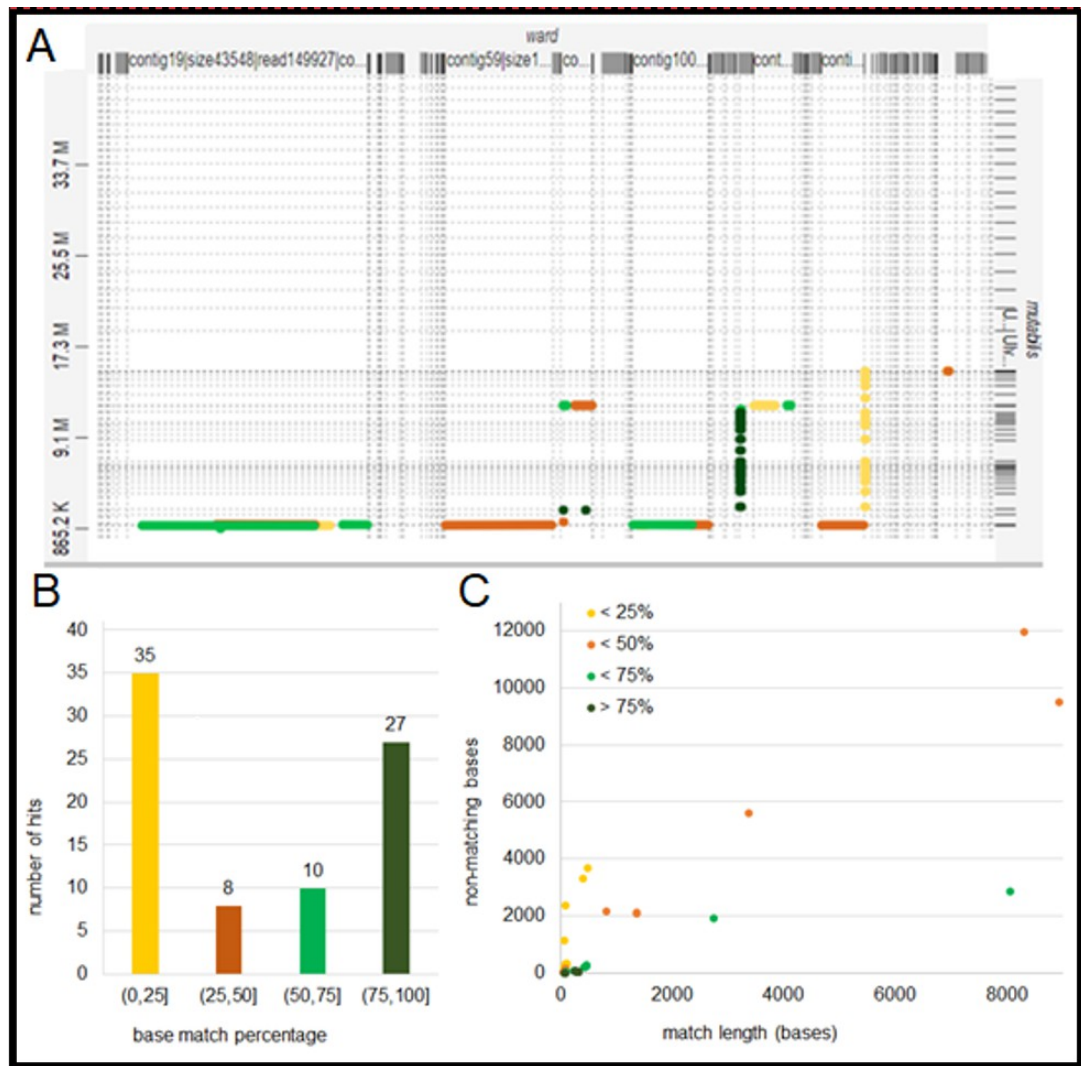


Figure 3.2 Dot plot analysis of query Ward and subject *U. mutabilis* genomes. Colors represent percent identity between the target and query and correspond to size classes presented by the dot plot software. Forest green, kelly green, orange, and yellow indicate matches of 75-100%, 50-75%, and 25-50%, and <25% identity, respectively. (A) Dot plot output. Pseudoassembly contigs on the horizontal axis are arranged in ascending numerical order. *U. mutabilis* contigs on the vertical axis are ordered such that the best resulting graphical output is achieved. Several *U. mutabilis* contigs have been cropped to focus on the entirety of Ward dot plot hits. No matches were omitted from the image, only unmatched *U. mutabilis* contigs (Cabanettes and Klopp 2018). (B) Size distribution histogram of matches in dot plot output. Numbers above bars are equal to count. (C) Scatterplot analysis of size class within each identity grouping against non-matching base counts. Unity line ($m=1$) overlays the x-axis.

Visual inspection of the dot plot identifies 6 specific Ward contigs (19, 26, 59, 100, 124, 162) which are of interest based on the unique arrangement of points on the dot plot. While sequence match length is not visually indicated by these graphical representations, this data is accessible via supplemental association files. Parsing of match location, length, identity, and mismatch base count from the PAF provide a

more representative summary of the dot plot data. Of hits between Ward and *U. mutabilis* greater than 25% identity, 60% fall within the range of 75-100% match. Of the total hits on the dot plot, including those of 25% match or less, these high identity hits represent 34% (Figure 3.2B). Of course, consideration of percent identity without match length is a weak analysis. These data indicate low identity matches (<50%) fall on the shorter end of the match length spectrum and are likely short, repeating sequences or sequences too short to represent a meaningful match or a recognizable gene hit in the reference genome. Generally, and as expected, higher identity matches present at longer match lengths and are assumed to be representative of truly homologous genomic regions (Figure 3.2C). The Dgenies PAF files is available (Digital Appendix 3.1).

Similarly, pseudoassembly Langston does not present in a typical dot plot graphical output. That is, notably missing from the figure is a characteristic diagonal line indicative of closely related species. Two patterns in particular appear in the comparison between Langston and *U. mutabilis*: a lack of low variability matches and few hits in the Langston pseudoassembly that recur in many contigs of the reference genome (Figure 3.3).

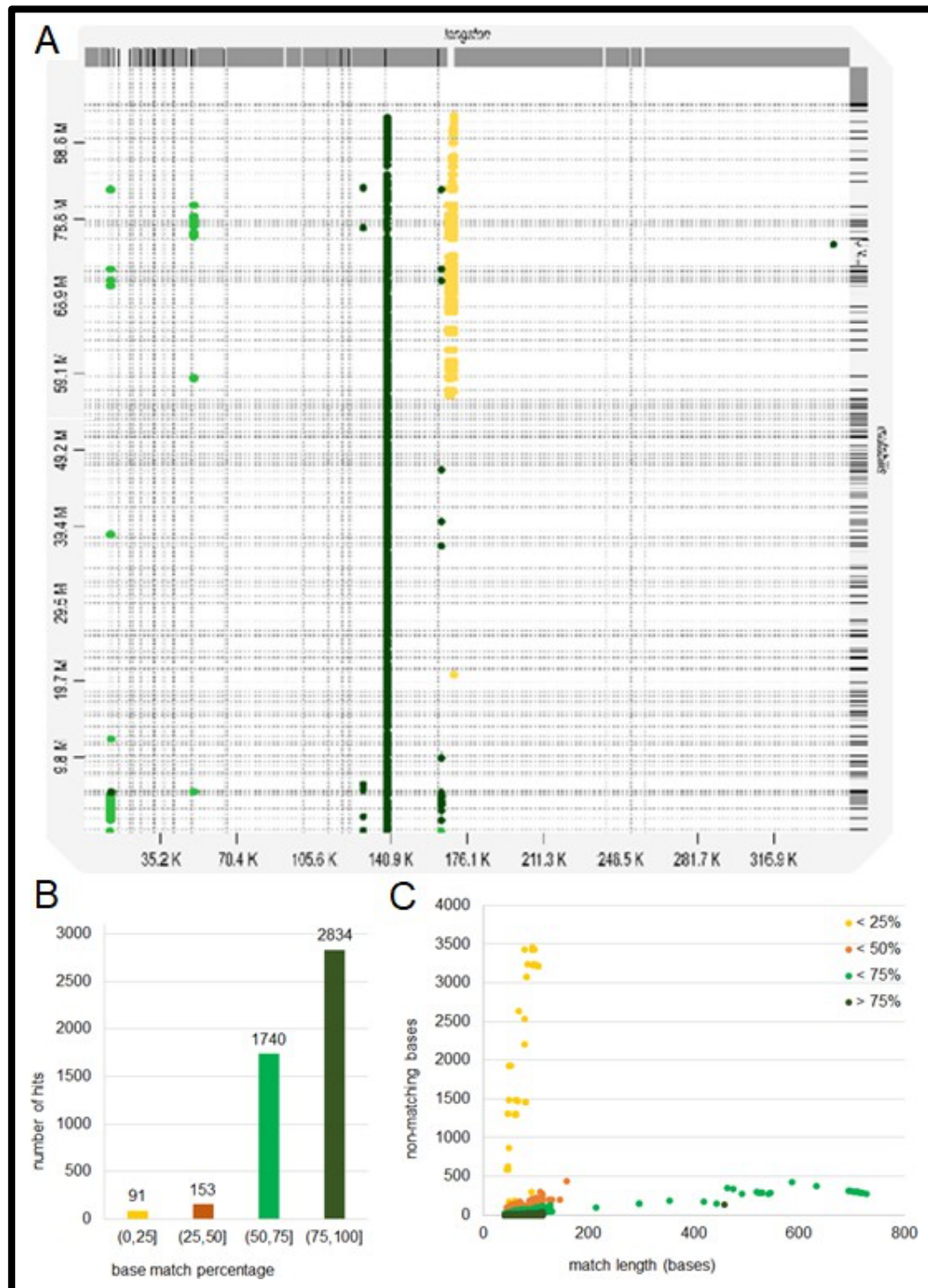


Figure 3.3 Dot plot analysis of query Langston and subject *U. mutabilis* genomes. Colors represent percent identity between the target and query and correspond to size classes presented by the dot plot software. Forest green, kelly green, orange and yellow indicate matches of 75-100%, 50-75%, 25-50%, and <25% identity, respectively. (A) Pseudoassembly contigs on the horizontal axis are arranged in ascending numerical order. *U. mutabilis* contigs on the vertical axis are ordered such that the best resulting graphical output is achieved (Cabanettes and Klopp 2018). (B) Size distribution histogram of matches in dot plot output. Numbers above bars are equal to count. (C) Scatterplot analysis of size class within each identity grouping against non-matching base counts. Unity line ($m=1$) overlays the x-axis.

Pseudoassembly sequences that we see repeat in many contigs in the reference genome are mainly from Langston contigs 38, 497, and 619. Two of the three (38 and 497) represent high identity matches. Uniquely, Langston contig 491 recurs in well over 90% of reference genome contigs (Figure 3.3A). This recurring match skews the percent identity match heavily towards the 75-100% bin, and results in nearly 120 times the number of matches as compared to the Ward analysis (Figure 3.3B, Figure 3.4). Even so, size distribution data indicates these high identity matches are very short in length (<200 bases). The longest matches from this analysis fall within the 50-75% match bin (Figure 3.3C). The Dgenies PAF files are available (Digital Appendix 3.2).

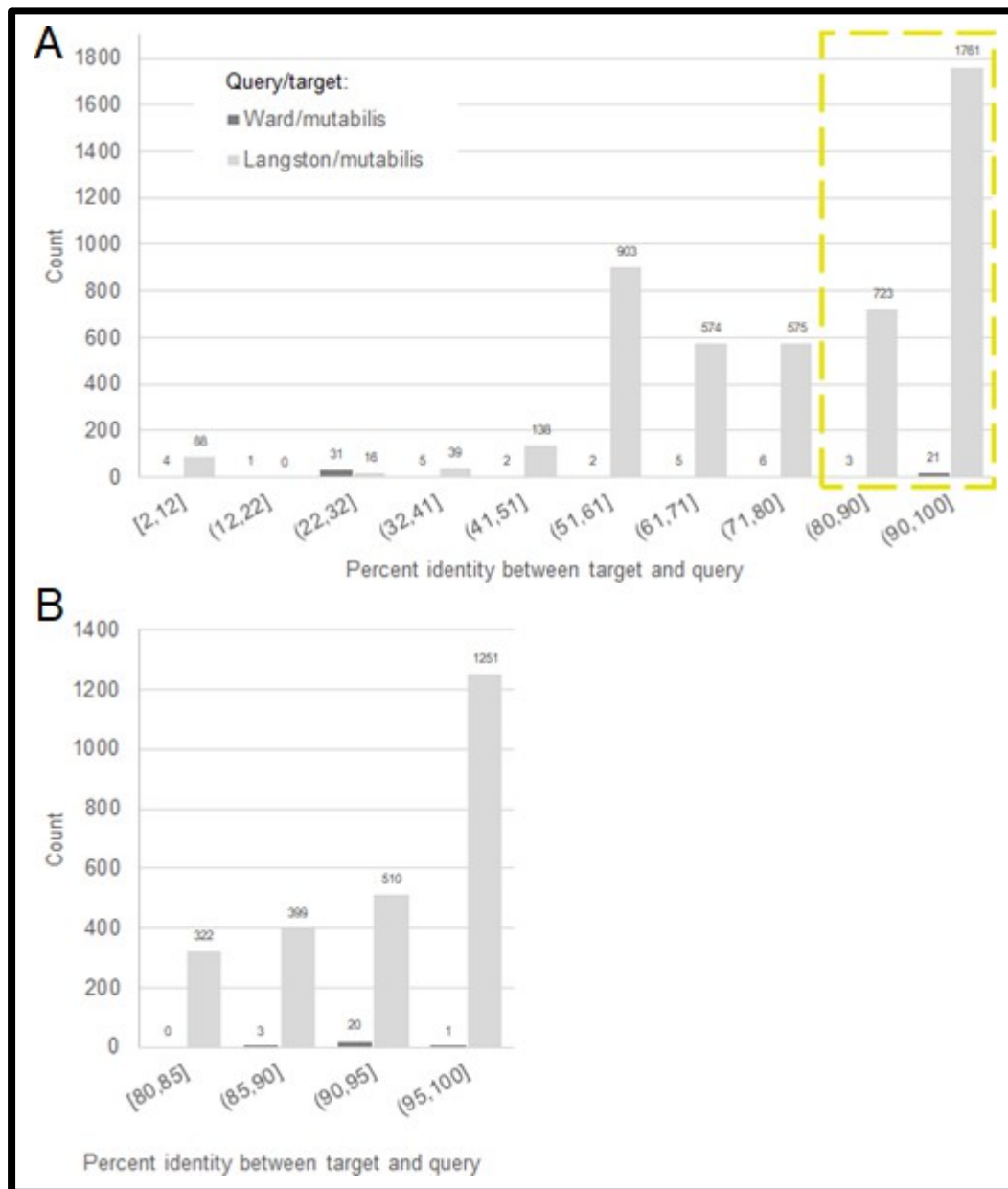


Figure 3.4 Dot plot data extraction of identity matches between each query and the target, *U. mutabilis*. Figure A shows the full range of percent matches. Figure B focuses on highest identity matches between 80 and 100%, selected in yellow from A.

Recalling the Ward (19, 26, 59, 100, 124, 162) and Langston (38, 497, 619) contigs which present unique dot plot findings, analyses were conducted to identify the content of the strongest hits and determine the origin of those unexpected findings.

Table 3.1 Percent variation accounted for along *U. mutabilis* contigs by pseudoassembly.

<i>U. mutabilis</i> contig	Nucleotide range (reference genome)	Ward / Langston hit count (% total variation)	% Variation in <i>U. mutabilis</i> contributed
178	1832 - 108884	12 (99.8%) / 4 (0.2%)	9
1	969403 - 3393024	2 (55.4%) / 4 (44.6%)	18
5	35499 - 1476650	1 (15%) / 4 (85%)	10
177	4015 - 17817	3 (100%) / -	11

A discontinuous megablast of the three assemblies identified four *U. mutabilis* contigs for which greater than two pseudoassembly contigs or singlets aligned with zero or 1 gap (Table 3.1). The percent variation along each stretch of *U. mutabilis* was attributed to each pseudoassembly. Not surprisingly, we see a range of scenarios when accounting for diversity along the reference genome sequences. Either the majority of the variation can be attributed to Ward or to Langston, or there is a relatively even split between the two. At the very least, this analysis provides evidence that the pseudoassemblies are not the same species, as three-way comparisons of diversity do not yield parallel results.

Table 3.2 Annotation of strongest query hits between pseudoassemblies and U. mutabilis reference genome.

	Query contig: sequence length (bases)	>200 base hits to target (mutabilis)	Annotations	Target genes and functions
Ward	59: 19592	3	1	<i>ATP binding, magnesium chelatase activity, chlorophyll biosynthetic process</i> UM178_0032.1
	161: 7851	6	1	<i>ATP binding, magnesium chelatase activity, chlorophyll biosynthetic process</i> UM178_0032.1
	208: 3443	3	1	<i>nucleus, phosphoric diester hydrolase activity, DNA repair</i> UM003_0084.1
Langston	139: 1035	100	15	<i>nucleic acid binding, DNA integration, zinc ion binding</i> UM004_0319.1 UM091_0009.1 UM110_0032.1 UM006_0089.1 UM95_0057.1 UM124_0010.1 UM024_0063.1 UM95_0059.1 UM147_0003.1 UM062_0005.1 UM104_0016.1 UM_156_0033.1 UM078_0045.1 UM110_0026.1 UM233_0001.1
	304: 1868	250	14	<i>nucleic acid binding, DNA integration, zinc ion binding</i> UM009_0059.1 UM095_0059.1 UM161_0001.1 UM039_0062.1 UM103_0002.1 UM163_0011.1 UM063_0017.1 UM110_0026.1 UM180_0005.1 UM076_0024.1 UM110_0032.1 UM233_0001.1 UM091_0009.1 UM124_0010.1
			2	<i>transcription factor activity, sequence specific DNA binding, catalytic activity, regulation of transcription, DNA-templated, kinase activity, phosphorylation</i> UM039_0014.1 UM163_0011.1
	619: 3976	99	15	<i>nucleic acid binding, DNA integration, zinc ion binding</i> UM004_0319.1 UM091_0009.1 UM110_0032.1 UM009_0059.1 UM095_0057.1 UM124_0010.1 UM024_0063.1 UM095_0059.1 UM147_0003.1 UM062_0005.1 UM104_0016.1 UM156_0033.1 UM078_0045.1 UM110_0026.1 UM233_0001.1
			1	<i>peptidyl-prolyl cis-trans isomerase activity, protein peptidyl-prolyl isomerization, protein folding</i> UM001_0068.1

High identity sequence stretches (megablast) with hits of 1000 bases or greater were extracted from each association table dot plot source data and aligned against the *U. mutabilis* reference genome (Table 3.2). Interestingly, contigs flagged in the dot plots as unique also presented in the subset of contigs and sequences matching these parameters. Ward hits to the reference genome were genes associated with ATP binding, magnesium chelatase activity, chlorophyll biosynthetic process, phosphoric diester hydrolase activity, and DNA repair. Langston hits to the reference genome were a bit more diverse (nucleic acid binding, DNA integration, zinc ion binding, transcription factor activity, sequence specific DNA binding, catalytic activity, regulation of transcription, DNA-templated, kinase activity, phosphorylation, peptidyl-prolyl cis-trans isomerase activity, protein peptidyl-prolyl isomerization, and protein folding).

Table 3.3 Ward and Langston BLAST results from non-matched *mutabilis* hits.

	Query Contig: sequence length (bases)	Count of >200 base hits to target (<i>mutabilis</i>)	Genera identification (number of species and evolutionary group)	Consensus function
Ward	19: 43548	11	<i>Ulva</i> : 17 other: 45 (green algae)	chloroplast genome
	69: 5717	4	<i>Ulva</i> : 17 other: 22 (green algae and mosses)	mitochondrial genome
	100: 13830	0	<i>Ulva</i> : 17 other: 45 (green algae)	chloroplast genome
	136: 7317	4	<i>Ulva</i> : 17 Other: 2 (green algae)	mitochondrial genome
Langston	50: 3381	0	<i>Ulva</i> : 2 Other: none	mating locus
	497: 1046	1	<i>Ulva</i> : none other: 31 (birds, rodents, and bony fishes)	artifact*

Considering the percentage of data unmatched to the reference genome, an investigation into “unmatched” sequences from each pseudoassembly was warranted. Few stretches of pseudoassembly sequence which met the aforementioned parameters returned no data from genes from the reference genome database. These examples, which include Ward 19 and 100 and Langston 497 (previously identified), were sent through a BLAST search to categorize their origin. In all but one case, the closest relatives were green algae or moss (Table 3.3). Most of these examples provide hits to close relative mitochondrial or chloroplast DNA, which explains why there is a notable lack of gene definition from the reference genome (Table 3.3). The *U. mutabilis* chloroplast and mitochondrial genomes are published separately, and therefore would not be expected to present hits based on a BLAST search (Melton et al. 2015, Liu et al. 2021). We can reasonably conclude that the pseudoassemblies contain not only genomic DNA, but also chloroplast and perhaps mitochondrial DNA.

A single outlier in this analysis, Langston 497, presents as a gene identified only in vertebrate taxa. This inconsistency is not assumed to be a true finding of the analysis and is presumed a sequencing artifact. Another contig of particular interest in the pseudoassemblies is Ward contig 19, not only because it is the longest contig resolved by the assembly pipeline, but also because we see it in the dot plot (Figure 3.2A) and in the unmatched hits (Table 3.3).

Chapter 4: CHARACTERIZATION OF NEWLY SEQUENCED CHLOROPHYTES

4.1 Introduction

In consideration of the trending nature of green algal whole genome sequencing, we saw opportunity to compare the *Ulva mutabilis* genome and the newly available *Ulva prolifera* and *Caulerpa lentillifera* genomes.

A recent paper published by He et al. (2020) in the Journal of Oceanology and Limnology introduces a draft genome for *Ulva prolifera*, a previously presumed and now confirmed close relative of *Ulva mutabilis*, which is used extensively in green algal research. Finding the demographics and characterization proposed in this paper as incomplete, here we propose a more comprehensive analysis of the composition of *Ulva prolifera* in the context of the widely accepted reference genome, *Ulva mutabilis*.

The *Caulerpa* genome, published by Arimoto et al. in 2019, provides a different context for the *Ulva* genome comparisons. An ulvophyte itself, the unique morphology of these organisms, coupled with their freshwater life history provides for an interesting comparison. *Caulerpa* clusters closely with *Ulva* in green algal lineage phylogenies, and thus, we expect to see some strong matches between both *U. mutabilis* and *U. prolifera*.

Specifically, there are morphological and compositional distinctions between these two ulvophyte groups. Sulfated cell wall constituents vary between *Ulva* and *Caulerpa*. The former presents uronic acid-containing polysaccharides while the latter contains uronic-limited polysaccharides. *Ulva* cell walls also contain markedly increased soluble ulvan and cellulose components as compared to *Caulerpa* (Domozych et al 2012). On a whole organism constituent level, *Ulva* contains higher concentrations of protein and fiber, while *Caulerpa* exceeds its cousin in lipid, carbohydrate, and moisture content (Ratana-Arporn and Chirapart 2006). Based on

lineage and varying life history strategy, the genome comparison of *Ulva* and *Caulerpa* is expected to yield overlap in metabolically essential gene functions, but potentially few others.

4.2 Methods

Genome demographics were collected from the publications in which they were introduced (De Clerck et al. 2018, He et al. 2021, Arimoto et al. 2019). Overall demographics were visualized by use of bandage plots (Wick et al. 2015).

Assembly data were analyzed using dot plot via the platform Dgenies (Cabanettes and Klopp 2018). Contigs were arranged according to best-fit, algorithmically by way of the central resolution line. Matching sequences between both target and query were omitted from the analysis if the sequence size fell between 0.001-0.2% of the dot plot width, a measure of strong precision hits, and stroke width was set to the maximum allowable to create the visuals presented here. Analysis of dot plot components were extracted from the corresponding association tables.

Annotation of selected sequences was obtained by extraction and alignment in Geneious Prime 2021.1.1 and annotation and search using NCBI BLAST and ORCAE (Sterck et al. 2012). These relationships were explored further using BBmap aligner with medium to low sensitivity processing. Any large-scale alignments were viewed in BLAST to identify genomic content and gene ontology. Further data processing was accomplished by manually extracting data from the dot plot PAF file in conjunction with select alignment comparisons, including ORF finder, in Geneious Prime. Details of each specific analyses is included alongside the results to illustrate the analytical approach with clarity and purpose.

4.3 Results

The sequencing of *Ulva prolifera* and *C. lentillifera* offers us virtually limitless options for novel comparisons among quality chlorophyte genome assemblies. Our approach to identifying the most pressing questions answerable in this massive dataset was driven by identifying unique patterns, beginning from the most basic genomic comparisons and allowing the data to drive the path forward. The first step in approaching the dataset was to define the general demographic parameters of each assembly, followed by a more careful dissection of general structure, alignment, and content. The beginning of this analysis reflects the processes developed with the novel pseudodassemblies, previously described (Chapters 2-3).

4.3.1 Comparative demographics

An overview of demographics provides high level comprehension of overall genomic organization. Here we present these demographics from the reference genome and the recently published *U. prolifera* and *C. lentillifera* (Table 4.1).

The *U. mutabilis* 98.5 Mbp genome is predicted to cover 98.5% of the haploid genome (de Clerck et al. 2018). Considering the *U. prolifera* assembly covers a predicted 75% of the haploid genome, it appears the recovery of 87.9 Mbp is an artifact of those inherent losses.

Table 4.1 Summary statistics and basic demographics of sequenced chlorophytes (adapted from De Clerck et al. 2018, Arimoto et al. in 2019, and He et al. 2020). NR = not reported. UK = unknown.

	<i>U. mutabilis</i>	<i>U. prolifera</i>	<i>C. lentillifera</i>
Genome size (Mbp)	98.5	87.9	28.65
Number of contigs	318	597	534
N50 (bases)	600,008	675,063	950,000
L50	54 (contig-based)	39 (contig-based)	14 (scaffold-based)
GC	57%	56%	40%
Number of genes	12,924	10,311	8,818
Repetitive elements (%) (reported repeat families)	35 (74)	22 (NR)	6.7 (UK)
Long terminal repeats (Mbp)	15.3	6.4	UK
Long interspersed elements (Mbp)	9.3	0.48	UK
Genome integrity (%)	98.5	75.3	86.4

As expected, the *Ulva* genus representatives present similarly compared to the more distantly related *Caulerpa* assembly. Considering approximately 14% of the *C. lentillifera* genome is unaccounted for in the assembly, this distant Ulvophyceae relative is, not unexpectedly, nearly one third the size its more morphologically complex cousins (Figure 4.1). This ratio, however, is not maintained when considering gene number, number of contigs, and N50. GC content in *Caulerpa* is quite low compared to *Ulva* sp. at 40%, although, this may possibly be an artifact of the ~14% of the genome which is missing from the assembly.

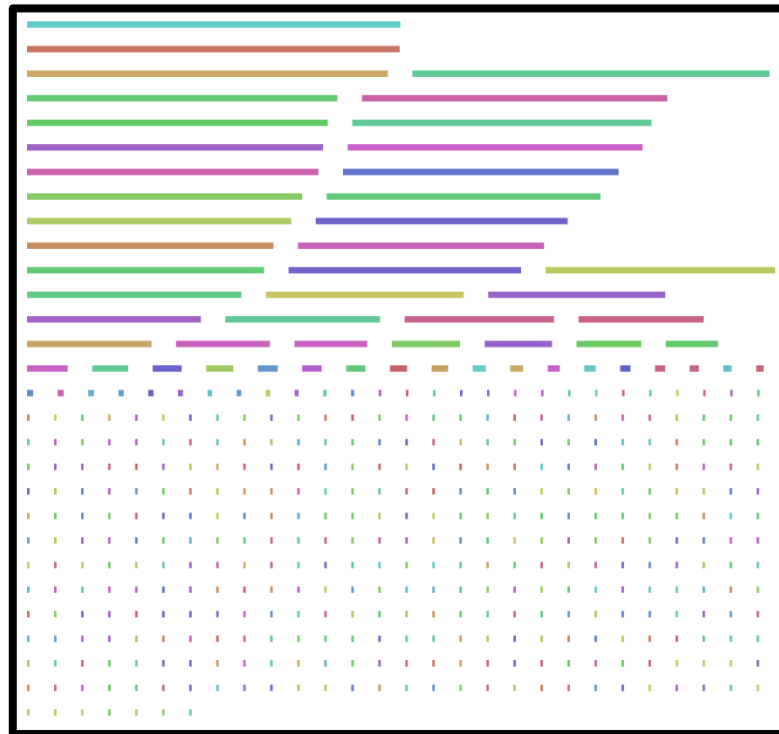


Figure 4.1 *Caulerpa lentillifera* bandage plot. The upper- and left-most turquoise band represents the length in bases of the longest scaffold in the genome; 1,292,516 bases (Wick et al. 2015).

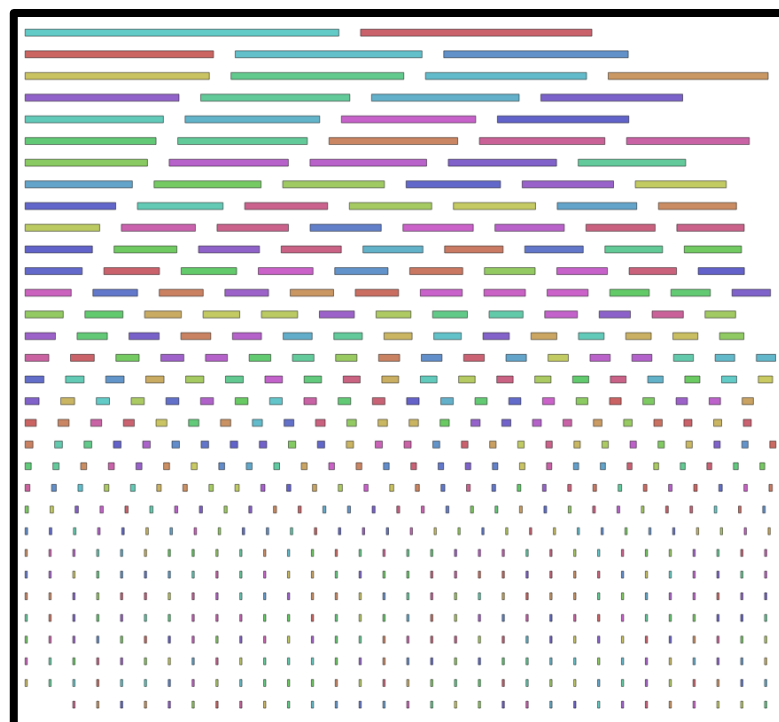


Figure 4.2 *Ulva prolifera* bandage plot. The upper- and left-most turquoise band represents the length in bases of the longest contig in the genome; 2,662,757 bases (Wick et al. 2015).

The longest contig in the reference genome represents ~4% of the overall genome length (Figure 4.1) while the longest contig in *U. prolifera* represents ~3% of the overall genome length (Figure 4.2). Logically, L50 increases because the whole genome size in bases also increases, therefore, it takes more contigs to hit the 50% genome base benchmark. For LTRs and LIEs genomic constitution decreases by 58% and 95%, respectively, between the reference genome and *U. prolifera*.

4.3.2 Multiple genomic comparison

The next step in this comparison involves taking a closer look at the content of the reference genome and that of each newly sequenced Chlorophyte genome. Much like the *De novo* genome pseudoassembly analyses, this process was data-driven and began with an overall view of homology via dot plot analysis, followed by investigation of interesting trends and unique characters presented in the figures (Cabanettes and Klopp 2018). Due to the relationship between each newly sequenced genome and the reference genome, there is a clear distinction in data availability favoring *U. prolifera* over *C. lentillifera*. There are, however, interesting findings to be discussed by comparison to both species and to both the reference nuclear and chloroplast genomes.

NUCLEAR GENOME ANALYSIS

Below are the results for the comparison of *C. lentillifera* with *U. mutabilis*, followed by the results for the comparison of *U. prolifera* with *U. mutabilis*. A final comparison of *U. prolifera* and *C. lentillifera* adds perspective to this trio but is of little analytical value due to a lack of substantial positive identifications. We can consider this comparison very similar to that of *U. mutabilis* and *C. lentillifera* with considerably less data available to contribute to that analysis due to the relative size and coverage of the subject genome (*U. prolifera*).

Caulerpa lentillifera and *Ulva mutabilis*

Although these two species are separated by an estimated 600-800 million years, the newly available *C. lentillifera* data is an attractive addition to this analysis (Draisma et al. 2014, Arimoto et al. 2019). On this basis, the number of hits between the genomes (85) is a near even split between low identity (<25%) and high identity (>75%) hits, presumably representative of short diverse sequences and highly conserved ones, respectively (Figure 4.3A-B). Here we present the unsorted dot plot version, with contigs (*U. mutabilis*) and scaffolds (*C. lentillifera*) ordered ascendingly. This presentation was selected based on the resolution of the sorted version and a lack of a central identity diagonal.

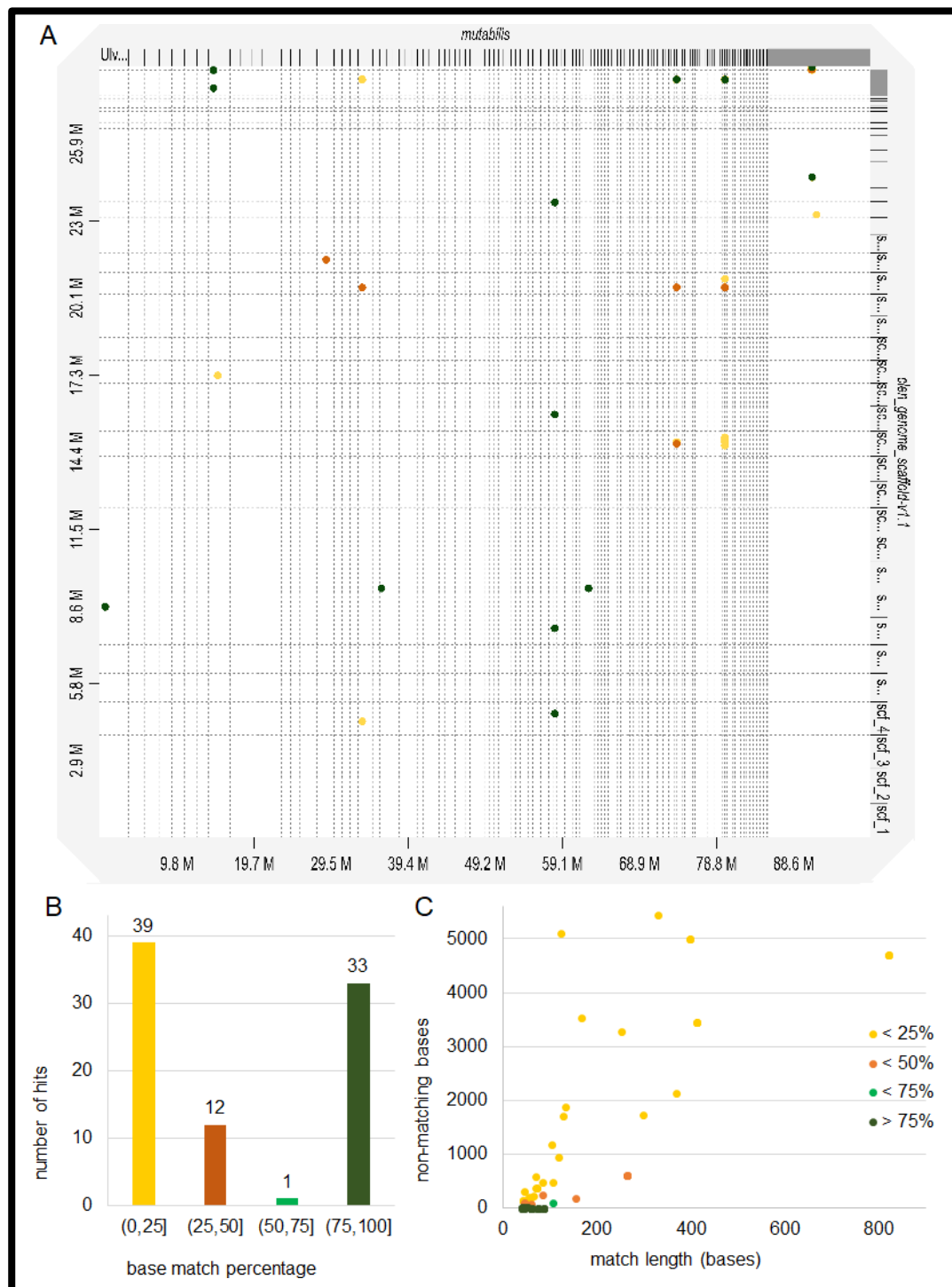


Figure 4.3 Dot plot analysis of query *C. lentillifera* and subject *U. mutabilis* genomes. Colors represent percent identity between the target and query and correspond to size classes presented by the dot plot software. Forest green, kelly green, orange, and yellow indicate matches of 75-100%, 50-75%, 25-50%, and <25% identity, respectively. (A) Dot plot output. Contigs on the horizontal and vertical axes are unsorted and appear in numerical order vertically towards the top axis and horizontally towards the left axis (Cabanettes and Klopp 2018). (B) Size distribution histogram of matches in dot plot output. Numbers above bars are equal to count. (C) Scatterplot analysis of size class within each identity grouping against non-matching base counts. Unity line ($m=1$) overlays the x-axis. Y-axis is non-matching bases in the matched sequence.

Interestingly, when parsing hits by total match length and the count of non-matching bases, we see that all highly conserved hits (>75%) are short (40-87 bases), but close to 100% identity (Figure 4.3C). The Dgenies PAF files is available (Digital Appendix 3.4).

Ulva prolifera and *Ulva mutabilis*

Due to the evolutionary closeness of these species, we expected to see quite an impressive dot plot display, and we were not disappointed with the resulting interactive image. This analysis yielded nearly 6,000 unique identity sequences shared to varying degrees of agreement between the two genomes (Figure 4.4A). Of those hits, the majority (79%) are of high diversity (<25%) (Figure 4.4B). The remaining sequences spread fairly evenly among the two midlevel diversity bins (25-50% = 9% and 50-75% = 5%) and the highest identity bin (75-100% = 7%). Uniquely, higher identity hits (>25%) are longer in length than the low identity counterparts (<25%), a strong indication that this data is truly representative of closely related species (Figure 4.4C). Additionally, differing from Figure 4.3, the identity diagonal is apparent, with varying degrees of identity along the midline.

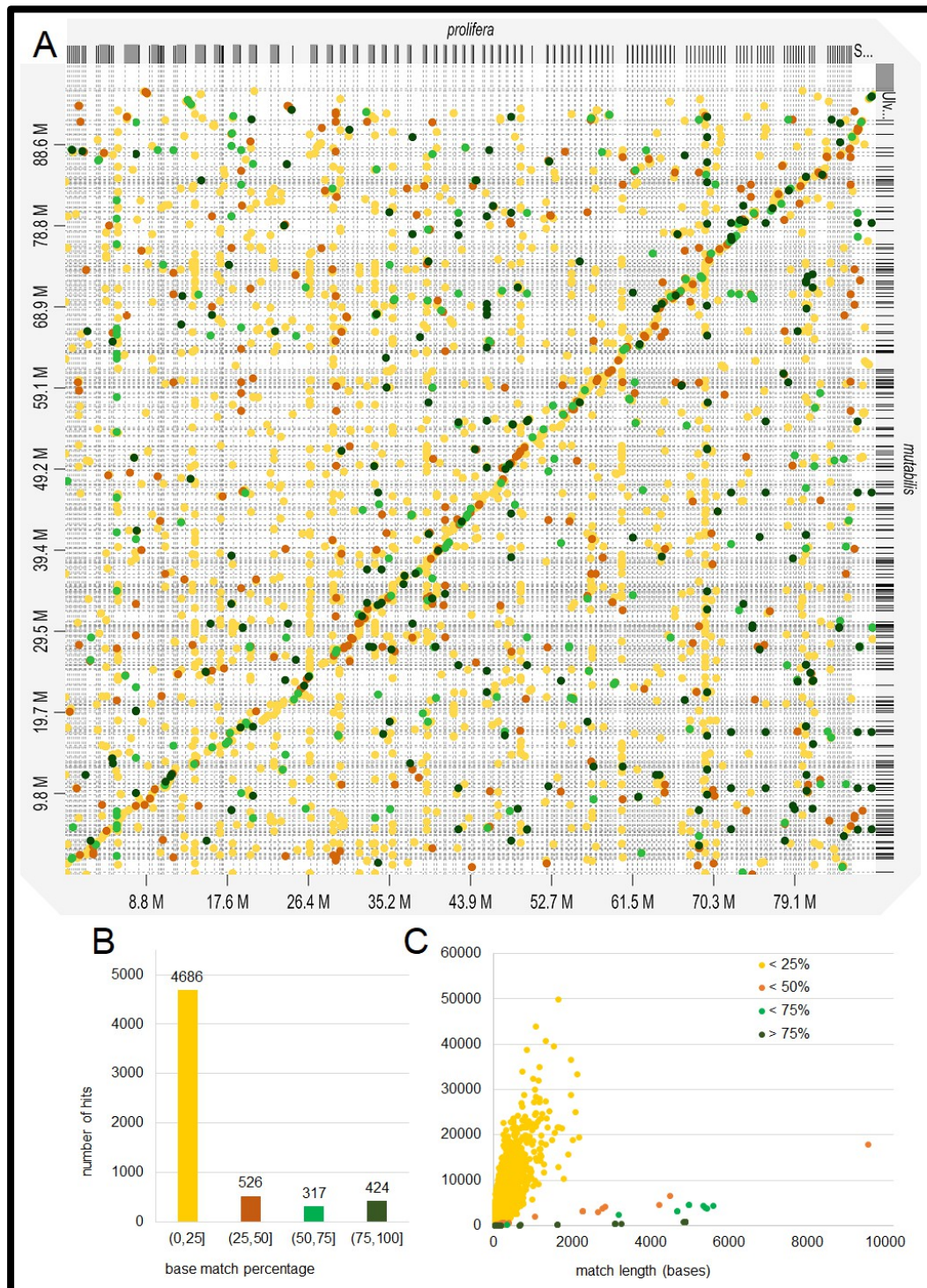


Figure 4.4 Dot plot analysis of query *U. proliferans* and subject *U. mutabilis* genomes. Colors represent percent identity between the target and query and correspond to size classes presented by the dot plot software. Forest green, kelly green, and yellow indicate matches of 75-100%, 50-75%, and 25-50% identity, respectively. (A) Dot plot output. Contigs on the horizontal and vertical axes are rearranged algorithmically based on best fit to the diagonal midline (Cabanettes and Klopp 2018). (B) Size distribution histogram of matches in dot plot output. Numbers above bars are equal to count. (C) Scatterplot analysis of size class within each identity grouping against non-matching base counts. Unity line ($m=1$) overlays the x-axis. Y-axis is non-matching bases in the matched sequence.

This comparison, indicative of two full size and nearly complete nuclear genomes, offers, essentially, limitless opportunities for comparative analyses. Given the vast amount of data available, four discrete features of the figure were identified as viable approaches for investigation (Dgenies PAF file, Digital Appendix 3.5); the first of which was a general description of content among *U. mutabilis* and *U. prolifera* matches. The second feature of the figure was low identity (<25%, yellow) continuous stretches of sequences which land on the central diagonal midline. These stretches of hits are interesting because although they fall in the highest diversity bin, the coordinates of their positioning within each genome are conserved. The third category of interest is high identity hits (>75%, forest green) which fall off the midline. In opposition to the previously described feature, these are near-identical sequences between the two genomes, indicating conservation of function but not structure, and therefore location, in the genome. Lastly, in consideration of the intricate detail encompassed by the dot plot, we explored interesting conformations of hits at high resolution. Essentially, interesting arrangements were identified by eye and explored theoretically based on the identity, location, orientation, and content of those hits.

Among the entire *U. mutabilis* and *U. prolifera* genomes, approximately 90% is unaccounted for in the dotplot. As previously described, the majority of the 10% accounted for are low identity hits. Therefore, we must assume those hits of high identity are physiologically or metabolically obligatory in function. To better visualize the relationship of these presumed conserved sequences, highest percent identity and longest alignment block length sequences were extracted, alongside *U. mutabilis* and *U. prolifera* contig identity (Table 4.2). The threshold for alignment block length was limited to matches over 1,000 bases in length, as this represents the presumable lower limit for gene length based on the *U. mutabilis* genome size and the total number of genes in the genome.

Table 4.2 High identity matches with the reference genome indicate conservation of genomic structure. Matches filtered by percent identity (>75%) and match length (>1,000 bases).

<i>U. mutabilis</i> contig	<i>U. prolifera</i> genome location	Percent identity (%)
Ulvmu_WT_24	SDUY01000076.1	88
Ulvmu_WT_24	SDUY01000076.1	88
Ulvmu_WT_24	SDUY01000076.1	88
Ulvmu_WT_24	SDUY01000076.1	86
Ulvmu_WT_24	SDUY01000076.1	86
Ulvmu_WT_24	SDUY01000076.1	86
Ulvmu_WT_24	SDUY01000076.1	86
Ulvmu_WT_99	SDUY01000076.1	86
Ulvmu_WT_99	SDUY01000076.1	86
Ulvmu_WT_99	SDUY01000076.1	86
Ulvmu_WT_99	SDUY01000076.1	85
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_120	SDUY01000076.1	86
Ulvmu_WT_178	SDUY01000494.1	88

There are several peculiarities apparent from this data extraction. First, there are very few hits after filtering by these parameters. These results are in agreement with those previously described (Figure 4.4C). Additionally, there are patterns of contig identity, most notably, 19 of the 20 sequences lie on a single contig in *U. prolifera*. Conversely, those hits are parsed into three contigs in *U. mutabilis*, however the overarching trend of general organization into discrete loci in the genome is conserved.

This led to an effort to identify the content within these hits. Interestingly, these sequences were aligning to a single location on each *U. mutabilis* contig and multiple positions along the single *U. prolifera* contig. Nearly 50% of alignments were identified as non-coding RNA large or small subunits (SSU) of ribosomal RNA, one alignment was an iron-stressed repressed RNA, and those remaining were EVIDENCEModeler (EVM) predictions for protein coding genes involved in transmembrane transport, metabolic processes (L-serine biosynthesis and oxidation-reduction), and homing endonucleases.

A second extraction represents mid-to-low range identity (10-58%) sequences parsed with the same alignment block length parameter as the previous analysis (Figure 4.4). This data set consists of 325 sequences between 10% and 59% identity.

Intuitively, this suggests the overwhelming majority of sequences in the original dot plot are less than 1,000 bases in length, comprising the omitted 5,607, or 94% of aligned sequences.

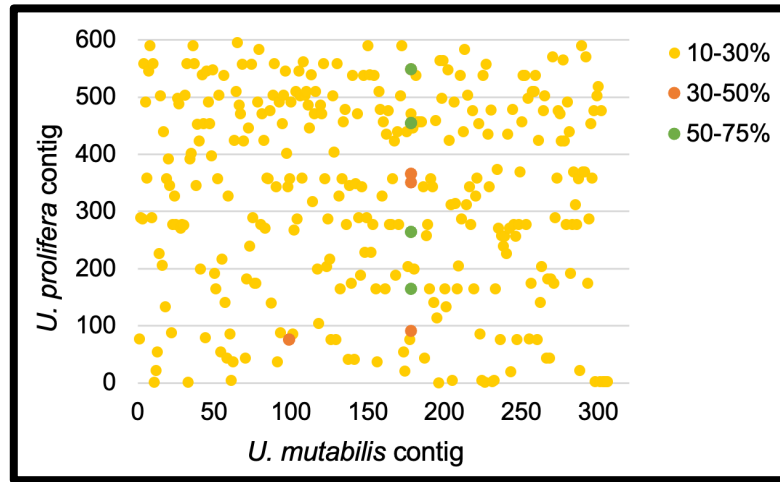


Figure 4.5 Coordinates of low to mid-range variation hits from *U. mutabilis* and *U. prolifera* dot plot. Color represents degree of variation.

Considering these aligned sequences fall on a grid, where the presence of a matched sequence is represented by a dot (Figure 4.5), the highest diversity points are relatively evenly distributed. This pattern is also visible in the algorithmically sorted version of the dot plot (Figure 4.4). There are, however, recurring patterns along the grid for the two stronger identity classes. The number of lowest identity aligned sequences is an order of magnitude larger than the other size classes. The mid-range identity hits centralize to a discrete number of *U. mutabilis* contigs among several *U. prolifera* contigs. Most notably, *U. mutabilis* contigs 99 and 178 recur in this dataset, reminiscent of the high identity matches present on *U. prolifera* contigs SDUY01000076.1 and SDUY01000494.1 (Table 4.2). In sharp contrast to that analysis, there is not a single point of origin from the query genome, but rather a relatively even disbursement of these hits among all *U. prolifera* contigs. Structure of more variable genomic regions is not conserved, as it appears to generally be with higher identity regions.

The next inquiry of the figure, low identity hits on the midline, is a bit more obscure. The identity and content of these hits are intriguing; *why is their positioning conserved but not the sequence, i.e., what is important about location above function?* Regions of the dot plot in which continuous stretches of low identity sequences were the primary composition of the central diagonal were sought. Several candidates were selected based on these criteria, with the additional caveat that the stretch of hits fell within one square of the dot plot, which is one single *U. mutabilis* contig and one single *U. prolifera* contig.

Once a stretch of hits was identified, the location (as match start base and match length) was extracted for each the subject and query. The resulting sequences were then extracted manually. There were between 11 and 20 for each of the four sample analyses with sequences match between 2,539 and 41,805 bases in length. As a proxy for sequence content, the open reading frame (ORF) count was determined between each individual sequence hit (dot) in the stretch (continuous yellow span on the dot plot midline).

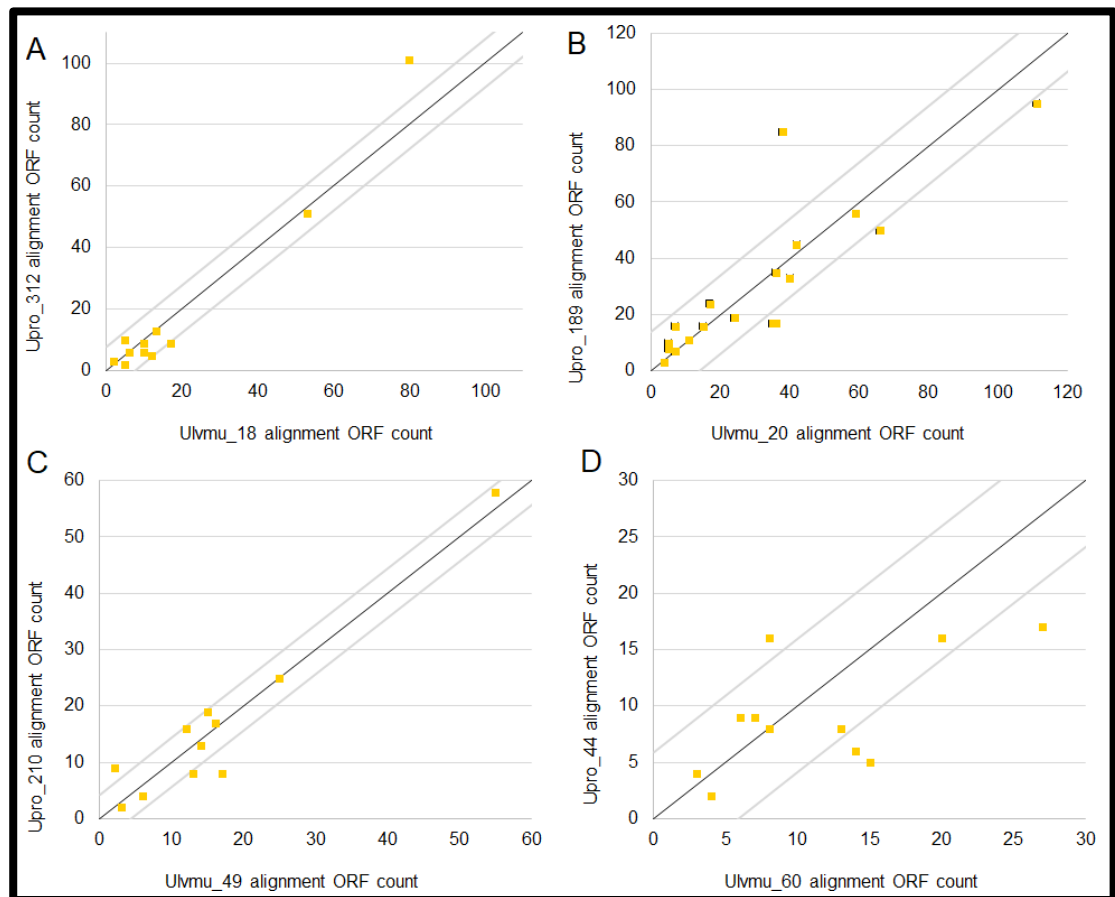


Figure 4.6 Comparison of open reading frame count between target and query, of which the dot plot hits are of low identity, fall along the dot plot midline, and are arranged in a continuous stretch within one single query contig and one single subject contig. Black line represents the unity line ($y = x$) and gray error lines represent the root mean square deviation (RMSD) of the data. Alignments A, B, C, and D correspond to specific dot plot regions (Digital Appendix 3.6).

The RMSD was plotted alongside the unity ORF count for each subject and query to find out of trend results, such as points which fall outside the limitations of the RMSD. Since the RMSD is determined by the spread of the data, data points which fall outside those parameters do not determine ORF count to be equivalent between the corresponding sequences in *U. mutabilis* and *U. proliferans* (Figure 4.6). In the specific case of ORF inequality between subject and query, the ratio was determined as way to predict duplication events. As it turns out, there is at least one of these events in each of the four specific cases investigated.

Table 4.3 Open reading frame count from low identity dot plot matches along the central diagonal.

Alignment contig (Ulvmu/Upro)	ORF count		Event classification in <i>U. prolifera</i>
	<i>U. mutabilis</i>	<i>U. prolifera</i>	
20 / 189	38	85	Duplication (2X)
20 / 189	36	17	Deletion (0.5X)
20 / 189	35	17	Deletion (0.5X)
18 / 312	17	9	Deletion (0.5X)
49 / 210	17	8	Deletion (0.5X)
60 / 44	15	5	Deletion (0.33X)
60 / 44	14	6	Deletion (0.5X)
18 / 312	12	5	Deletion (0.5X)
60 / 44	8	16	Duplication (2X)
49 / 210	2	9	Triplication (3X)

Further investigation of the points which fall outside of the RMSD of each dataset yield several out of trend hits, multiple from each target and query pairing, which specifically indicate a proportional change in ORF number (Table 4.3). By estimating the ratio of *U. mutabilis* to *U. prolifera* ORF count, we can determine what kind of expansion of deletion event has taken place in *U. prolifera*.

Of course, specifying the content of these sequences is another important measure in qualifying these results. To do so, the Geneious Prime EMBOSS tfscan function was applied to each sequence to identify transcription factors present within the previously established open reading frames. With regards to transcription factor identity, there are several which recur in each *U. mutabilis/U. prolifera* pairing and in several individual sequence alignments (NIT2, TAF-1, Opaque-2, CPC1, and Zmhox1a). Often, but not in every case, the number of transcription factors increase proportionally to the classification event.

A fourth and final point of interest in the dot plot (Figure 4.4) were high identity alignments (>75%, forest green), which fell off the midline. These hits are interesting because unlike the previous analysis, their sequence is highly conserved, but their location is not. Inherently, there is a huge opportunity for bias here, as the selection of hits chosen off the midline may be inadvertently selected based on a visual pattern. To account for this factor, a random number generator was used to pair one *U. mutabilis* contig to one *U. prolifera* contig. The probability of encountering a midline high identity match for any given *U. prolifera* contig is 0.0031%, and likewise for any given *U. mutabilis* contig is 0.0016%, so the random number generator would provide coordinates on the grid for which to select high identity hits where the likelihood of selecting a unity line hit was very rare. This process was repeated 138 times. Alignment length, including stop and start codon, match length, and percent identity were collected for both the *U. mutabilis* sequence and the *U. prolifera* sequence of the generated coordinate. If there were multiple high identity hits at any given coordinate, they were all included in the dataset so as to minimize bias. These matches were ordered and the cutoff for length was again set at 1,000 bases. Of these high identity hits positioned off the dot plot midline, 13 were characterized as probably candidates for containing large segments of potential *Ulva* genes (3-6Kb). The percent identity of the sorted subset was 86-88%. Finally, the *U. mutabilis* match sequence was extracted via ORCAE and any ontology predictions were included.

These hits spanned 12 genes in the *U. mutabilis* genome, all of which returned ontology for non-coding RNA, either large subunit or small subunit ribosomal RNA or iron-stressed repressed RNA. Two protein coding genes were also encompassed in this data set, coding for intron-encoded homing endonuclease and a small subunit ribosomal RNA gene.

Based on our previous findings regarding the presence of chloroplast DNA in the *U. mutabilis* reference genome and a newly published paper comparing the chloroplast genome between several *Ulva* species and individuals, a final analysis was conducted to compare the presence or absence of specific chloroplast genome characters between the reference genome and the newly published *U. prolifera* genome, with particular attention to genome organization and gene order (Liu and Melton 2021, He et al. 2020). Liu and Melton focus on seven *U. mutabilis* individual chloroplast genomes and a single *U. prolifera* chloroplast genome (2021). Arrangement and order are visualized using blocks corresponding to open reading frames, transfer RNAs (tRNA), ribosomal RNAs (rRNA), and core protein coding genes with chloroplast-specific functions, where blocks represent gene positioning along a hypothetical linear genome.

Table 4.4 Basic statistics of *U. mutabilis* and *U. prolifera* chloroplast genomes. Adapted from Xia et al. 2021 and Liu and Melton et al. 2021.

	<i>U. mutabilis</i>	<i>U. prolifera</i>
Size (bases)	94,226	93,066
Protein coding genes	63	66
tRNAs	23	26
rRNAs	4	3
A %	37.2	38.0
T %	37.0	37.0
G %	12.7	12.3
C %	13.1	12.5

The comparison of basic demographics in the two target species suggests they are remarkably similar, in terms of overall size, the number of coded genes, and base

(Table 4.4). The species seem almost indistinguishable based on these statistics alone, and this makes the level of variation we see in the dot plot all the more interesting. Intuitively, composition must account for the variability observed both within and between species. In an attempt to summarize characteristic differences in chloroplast genome composition and structure, the available data by Liu and Melton was extracted, aligned, and condensed against that from Xia et al. (2021).

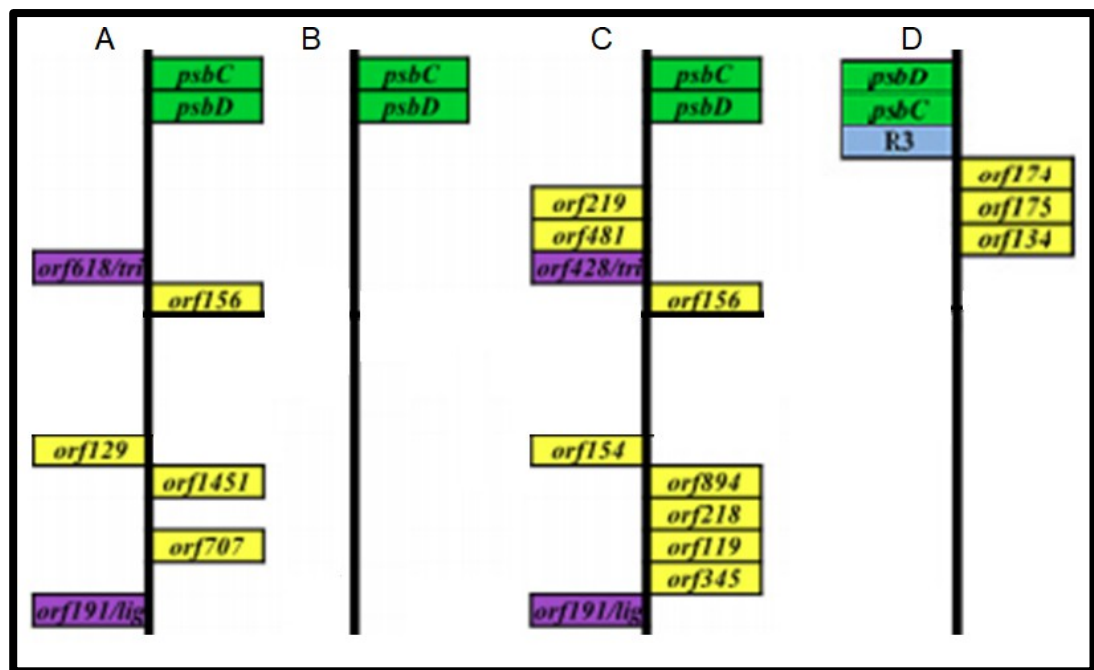


Figure 4.7 Comparison of chloroplast genomes between (A) *Ulva compressa* (*mutabilis*), (B) *Ulva compressa* (*mutabilis*) 2-3 and 5-7, (C) *Ulva compressa* (*mutabilis*) 4, and (D) *Ulva prolifera* (adapted from Liu and Melton 2021).

What resulted was a visualization of the key differences intraspecifically and between *U. mutabilis* and *U. prolifera* (Figure 4.7). The most prominent differences found were the order of photosystem II protein coding genes, inclusion of specific tRNA, and the positioning and order of ORFs. It is important to note that extraction of ORFs was performed using ORF finder in Geneious Prime both in this analysis and in that of the Liu and Melton. Additionally, identification of ORFs and extraction of sequences for content comparison was conducted in the published

genomes for both *U. mutabilis* and *U. prolifera*, and not in contig 178 of the reference genome, deemed in this thesis to partially cover the chloroplast genome (Ch 3).

The order of photosystem II protein coding genes *psbC* and *psbD* were reversed in *U. prolifera* as compared to the seven *U. mutabilis* counterparts. There were several ORFs which were not present in *U. prolifera*, and the ORFs in *U. mutabilis* seemed to assume a high degree of variability between individuals. More specifically, four characters were found to be exclusive to *U. prolifera*: R3 (a tRNA) and three ORFs (174, 175, and 134). Conversely, several ORFs were distinctly absent in the *U. prolifera* dataset but were present in at least one *U. mutabilis* genome (ORF 219, 418, 156, 154, 119, 218, 345, 191/lig, and 129). To verify the absence of these ORFs, the sequence was extracted in Geneious Prime based on the ORF finder results, aligned against the other genome, and run via BLAST against chlorophyte chloroplast genomes, with the goal of finding overlooked omissions or other out of trend results. Presumably, sequences from one chloroplast genome should provide hits to others based on the little variability we observe in the general demographics of these two genomes.

Not surprisingly, all three sequences from the ORFs in *U. prolifera*, which were not present in the three unique genome structures presented for *U. mutabilis*, were not found anywhere in the *mutabilis* genome. They also did not provide hits to any other chlorophyte genomes. For the list of ORFs which were present in one or more *U. mutabilis* individuals, but were absent from the *U. prolifera* genome, there are three distinct scenarios: (1) *U. mutabilis* ORFs which map to *U. prolifera* with high identity, (2) *U. mutabilis* ORFs which do not map to *U. prolifera*, nor do they map to other chlorophyte chloroplast genomes, and (3) *U. mutabilis* ORFs which do not map

to *U. prolifera*, but do map to other ulvophyte chloroplast genomes. In one particularly interesting example, ORF 191/lig from *U. mutabilis* does not map to *U. prolifera*, but maps to other chloroplast genomes, including *Caulerpa*.

Chapter 5: ORTHOLOGOUS GROUPS IN SEQUENCED CHLOROPHYTES: EVOLUTION OF GREEN SULFOTRANSFERASES

5.1 Introduction

Identification of orthogroups and orthologs between species has become a popular analysis in the past two decades as an alternative approach to more tedious methods of quantifying homology between similar sequences (Chen et al. 2006, Trachana et al. 2011, Altenhoff et al. 2011, Waterhouse et al. 2013). Traditional methods perform iterative comparisons of homologous sequences to estimate divergence but may not consider confounding factors such as variable gene evolution rates. OrthoFinder is an open-access statistical and comparative tool to analyze divergence groupings via phylogenetic trees of gene groupings, which accounts for traditional misrepresentations by analyzing whole genome orthology and duplication events to identify a rooted species phylogeny (Emms and Kelly 2015, 2019).

A crude OrthoFinder analysis following the multiple genome comparison previously described (Chapter 4) with the most recent proteomes of *U. mutabilis*, *C. lentillifera*, *Chlamydomonas reinhardtii*, and *Volvox carteri* yielded a widely familiar phylogenetic tree (Figure 5.1).

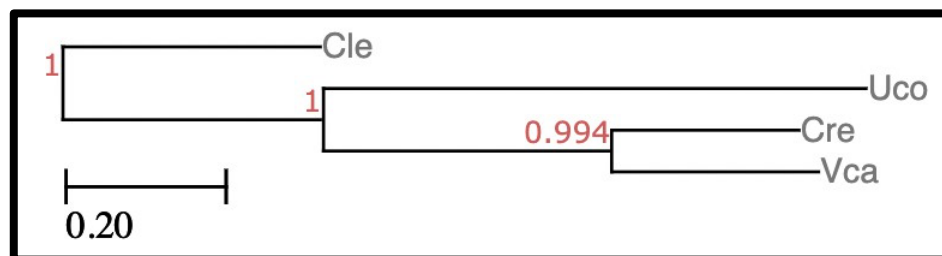


Figure 5.1 Phylogeny of four key green algal species based on proteome data.

This orthogroup analysis revealed a major expansion in membrane-bound sulfotransferases (SOT) in the green seaweeds. As their name suggests, the sulfotransferases catalyse the transfer of sulfate to substrates, using 3'-

phosphoadenosine-5'-phosphosulfate (PAPS) as the sulfate donor (Figure 5.2). There are two major classes of SOT, which contain similar structural features (Kakuta et al. 1997).

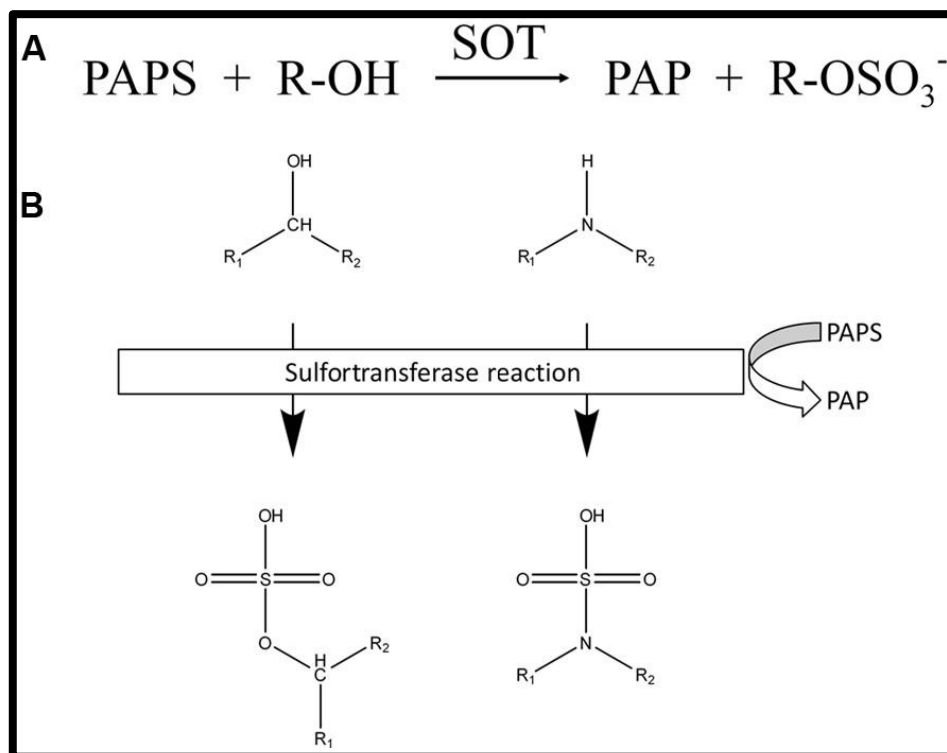


Figure 5.2 Catalytic reaction with sulfotransferases (SOT) showing products and substrates in the presence of 3'-phosphoadenosine 5'-phosphosulfate (PAPS) (adapted from Hirschmann et al. 2014).

At one side of one end of this β -barrel, highly conserved amino acid residues act to bind the sulfate donor, PAPS. At the other side of the same end, α -helices help to form a substrate binding cleft (Kakuta al. 1998). Many sulfotransferases have dimerization motifs (Petrotchenko et al. 2001). Sulfotransferases catalyse the addition of sulfate to a broad range of substrates with, in general, membrane-bound sulfotransferases acting within the Golgi on carbohydrates or glycoproteins and cytosolic sulfotransferases, which renders them more water soluble than other non-sulfated counterparts (Weinshilboum and Otterness 1994).

Table 5.1 Comparison of cell wall components and functional groups across major green algal groups (adapted from Paliwal et al. 2018).

Division (common name)	Cell wall	Functional groups
Chlorophyta (green algae)	Cellulose, hydroxyproline, glucosides (xylans and mannans)	Carboxyl and sulfate
Phaeophyta (brown algae)	Cellulose, alginic acid, sulfated mucopolysaccharides (fucoidan)	Carboxyl, amino, and sulfate
Rhodophyta (red algae)	Cellulose, xylans, several sulfated polysaccharides (galactans)	Carboxyl and sulfate

In the chlorophyte macroalgae, cell walls are sulfated (Cosenza et al. 2017, Ciancia et al. 2020) (Table 5.1). Previous analysis of *Ulva* cell wall constituents with Fourier-transform infrared spectroscopy (FTIR) have shown most sulfate residues reside in the pectin fraction, indicated by the peak at approximately 1230 cm⁻¹ (Figure 5.3). We know peaks at this particular wavenumber, in combination with subsidiary peaks at 650 cm⁻¹, represent sulfate compounds (Mainreck et. al 2011). Furthermore, Carbon-Hydrogen-Nitrogen-Sulfur (CHNS) analyses of *Ulva* reveal these samples to be Sulfur-rich.

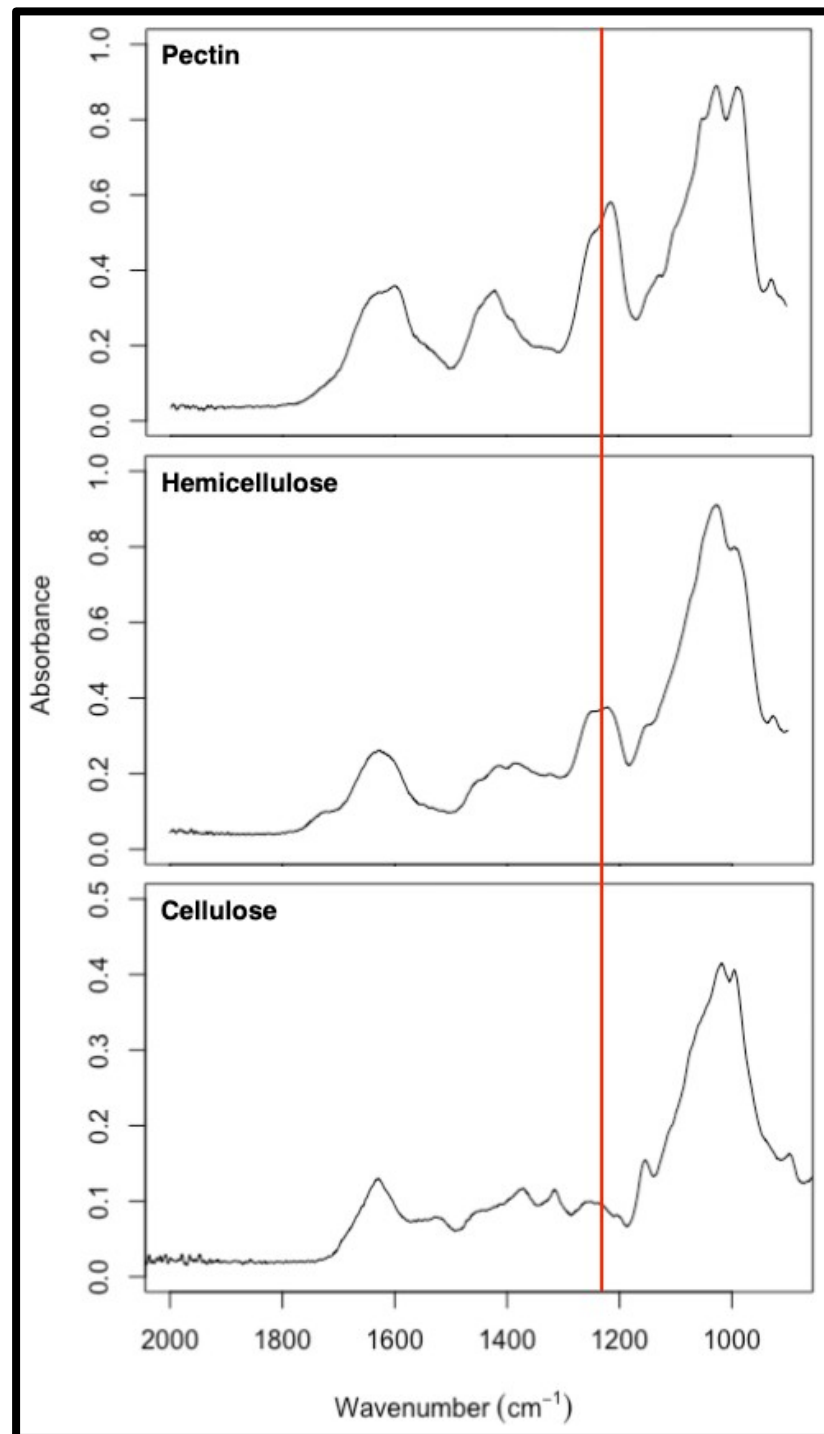


Figure 5.3 FTIR spectroscopy of *Ulva* cell wall pectin (upper), hemicellulose (middle), and cellulose (lower) fractions. Red line highlights wavenumber $\sim 1230\text{ cm}^{-1}$.

Marine sulfate levels have risen in several steps since the dawn of life, with the most recent being around 500 MYA (Canfield and Farquhar 2009). Because this postdates divergence of the green seaweeds, we test here the hypothesis that sulfotransferases evolved independently in the red and green seaweeds through separate horizontal gene

transfer (HGT) events.

Another interesting hypothesis postulates levels of cell wall sulfation in green algae may be a mechanism to cope with environmental stressors, such as salinity or dessication-rehydration cycles, (Kloareg and Quatrano 1988, Aquino et al. 2005, 2011, Pomin and Mourao 2008, Holzinger et al. 2015, Vishwakarma et al. 2019), although there is also evidence to the contrary (Arata et al. 2017). This theory fits with our genomic comparison of the distant relatives *U. mutabilis*, the sea-faring species, and *C. lentillifera*, the freshwater one. In this case, habitat is a proxy for molecular and physical stress. Here this relationship is explored in the context of comparative cell wall sulfation levels and comparison of orthologous SOT genes among diverse chlorophytes.

5.2 Methods

5.2.1 Total cell wall polymer extraction

Total cell wall extraction was carried out using standard alcohol-insoluble residue methods: thalli were frozen in liquid nitrogen and ground to a fine powder, still under liquid nitrogen. The powdered sample was washed twice more to remove pigments, in 96% and 100% ethanol respectively; after each wash the sample was centrifuged for 15 minutes at 3,000 g and the supernatant was discarded. The remaining pellet was washed twice in 2:3 methanol and chloroform solution to remove polar and nonpolar metabolites; each wash ran overnight with shaking and was followed by 15 minutes centrifugation at 3,000 g with the supernatant being discarded. Finally, the pellet was washed three more times in 100%, 65%, and 80% ethanol solutions; after each wash the suspension was centrifuged for 15 minutes and the supernatant was discarded. The remaining pellet was resuspended in 100% ethanol and dried to leave cell wall powder.

5.2.2 Extraction, separation, and analysis of cell wall fractions

To compare cell wall sulfation levels between a marine representative (*Ulva*) and a freshwater one (*Caulerpa*), fractionation of the total cell wall polymer sample was performed to isolate the pectin, alpha-cellulose, and hemicellulose extracts (Figure 5.4).

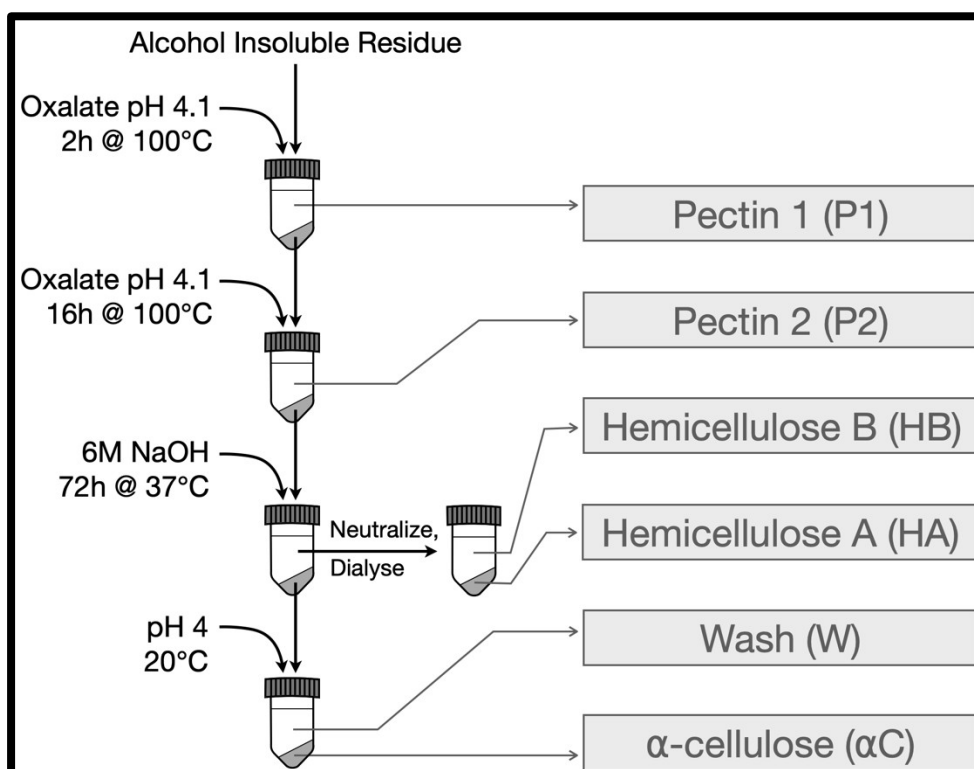


Figure 5.4 Extraction protocol for cell wall fractions (adapted from Mainreck et. Al 2011).

Cell wall fractions were extracted and separated using a modification of our previous method (O'Rourke et al. 2015). The 'pectin-like' fractions were solubilized from cell wall powder with 0.2 M ammonium oxalate (pH 4.0-4.3), at 100°C, for 2 hours, followed by an additional 16 hours. The 2-hour and 16-hour pectic extracts (P1 and P2) were dialysed and freeze-dried.

In some experiments, hemicellulose was then extracted from the oxalate-insoluble residue in 6 M NaOH, at 37°C for 72 hours. The extract was neutralized with acetic

acid and dialysed against water. Precipitated material from these operations (hemicellulose a, HA) was sedimented by centrifugation at 5000 g for 10 minutes, rinsed in water, and freeze-dried. The remaining solution (hemicellulose b, HB) was also freeze-dried. The NaOH in-extractable material was rinsed several times with pH 4 buffer, and the pooled washings (wash, W) were dialysed and freeze-dried. The final residue (a-cellulose, aC) was rinsed in water and freeze-dried.

Fourier-transform infrared (FTIR) spectroscopy was carried out on a PerkinElmer Frontier IR/NIR spectrometer equipped with UATR polarization and operating with a gauge force of 60 N.

5.2.3 *Phylogenetic analysis*

The proteomes for 12 sequenced algal species were downloaded from ORCAE, Phytozome (Goodstein et al. 2012), or EnSEMBL (Howe et al. 2021). A species tree was specified and used to direct OrthoFinder (Figure 5.5). The predicted proteomes of these algal species were parsed for protein descriptors containing the keyword ‘sulfotransferase’ (Table 5.2). All hits were then passed to InterProScan (Finn et al. 2017). Sequences matching this description which also had predicted transmembrane domains and the characteristic IPR037359 domain were tentatively assigned as membrane-bound sulfotransferases, while those that lacked transmembrane helices but had the cytoplasmic SOT IPR000863 domain were tentatively assigned as cytosolic sulfotransferases.

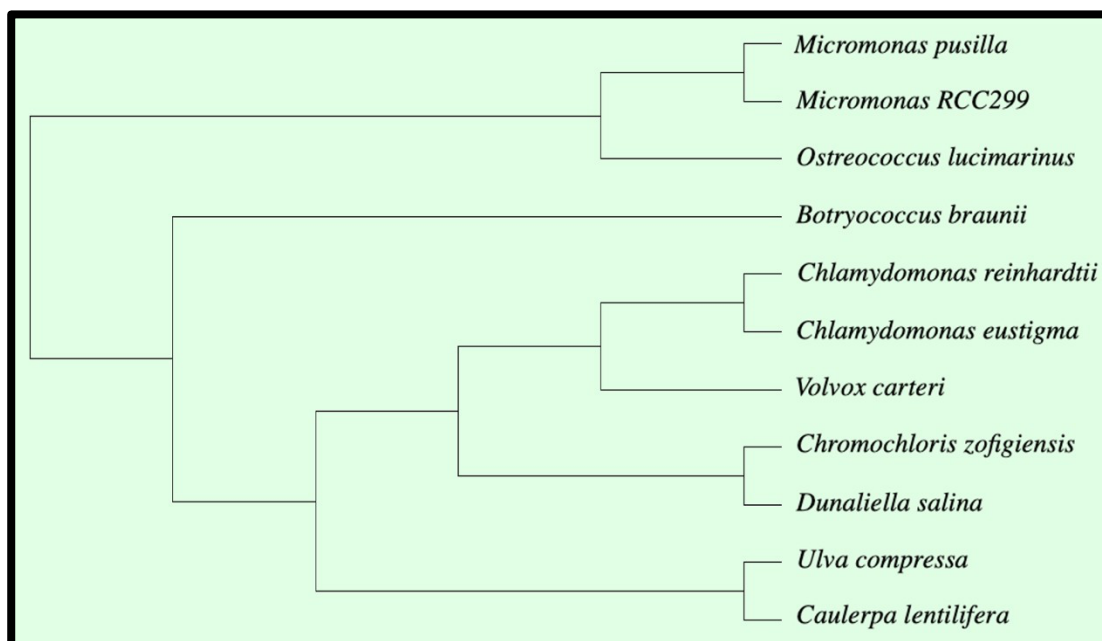


Figure 5.5 Phylogeny of proteomes used in Orthofinder analysis.

Reciprocal BLASTP searches were conducted for homologs to representative membrane sulfotransferases (UM075_0045 in *U. compressa*, XP_005719308 in *Chondrus crispus*, and g1228 in *C. lentillifera*), keeping any sequences with bit scores >50. PSI-BLAST was used to tentatively confirm SOT relationships for some sequences listed as hypothetical proteins. Multiple alignments were generated in Jalview using full-length sequences in Mafft. The evolutionary history was inferred using the Maximum Likelihood method running in MEGA X (Kumar et al. 2018), with partial deletion above 80%.

5.3 Results

5.3.1 General metabolism and proteomics

Analysis of the published proteome data resulted in data of predicted proteome size with the associated number of CHSTs (carbohydrate sulfotransferase), along with general descriptions of cell wall composition and details of habitat and other unique identifiers in select cases (Figure 5.5, Table 5.2).

Table 5.2 Proteome data and cell wall components of red and green micro- and macroalgae. ND = not determined. NA = not available.

Species	Predicted proteome size	Number of CHSTs	Cell wall	Notes
<i>Botryococcus braunii</i> (Chlorophyceae: Trebouxiophyceae)	20,284	0	Yes	Lakes and estuaries
<i>Caulerpa lentillifera</i> (Chlorophyceae: Ulvophyceae)	9,311	6	Sulfated carbohydrate (xylan-rich)	-
<i>Chlamydomonas eustigma</i> (Chlorophyceae: Volvocales)	14,105	8	Glycoprotein	Found in sulfur-rich mine water
<i>Chlamydomonas reinhardtii</i> (Chlorophyceae: Volvocales)	17,741	4	Glycoprotein	-
<i>Chondrus crispus</i> (Rhodophyceae: Florideophyceae)	9,606	1	Sulfated polysaccharides	-
<i>Chromochloris zofigiensis</i> (Chlorophyceae: Volvocales)	15,274	0	NA	Freshwater (of biofuel interest)
<i>Dunaliella salina</i> (Chlorophyceae: Volvocales)	16,697	0	NA	-
<i>Micromonas pusilla</i> (Chlorophyceae: Mamelliales)	10,660	0	None	
<i>Micromonas RCC299</i> (Chlorophyceae: Mamelliales)	10,103	0	None	
<i>Ostreococcus lucimarinus</i> (Chlorophyceae: Mamelliales)	7,796	0	None	-
<i>Volvox carteri</i> (Chlorophyceae: Volvocales)	14,247	ND	Glycoprotein	Freshwater
<i>Ulva compressa</i> (Chlorophyceae: Ulvophyceae)	11,000	18	Sulfated carbohydrate	-

Based on the previously described analysis (5.2.3), the green algae appear to have sulfatases and homologs to PAPS-producing enzymes (Table 5.2). Logically, species with no cell wall lack CHSTs (*C. zofigiensis*, *D. salina*, *M. pusilla*, *M. RCC299*, and *O. lucimarinus*).

5.3.2 SOT patterns in the green lineage

Predicted proteome size does not appear to be a good indicator of the CHST count per genome (Figure 5.6). There are, however, a few interesting trends that arise from this comparison. First, proteome size, and perhaps therefore genome size, is generally larger in microalgae. Additionally, there seems to be a mid-size range which produces proteomes with the highest number of CHSTs. It appears this size range is in the low 10,000 count, for a maximum CHST count of 18 (*U. compressa*) in this specific dataset. Again, those data which fall on the x-axis are not considered towards this estimation, since they would be expected to lack CHSTs based on morphology and/or cell wall composition. Compared to the other datapoints, there is an expansion in *U. compressa* (most vertical red point) and *C. eustigma* (most vertical blue point), compared to the rest of the species for each size class (macroalgae and microalgae, respectively).

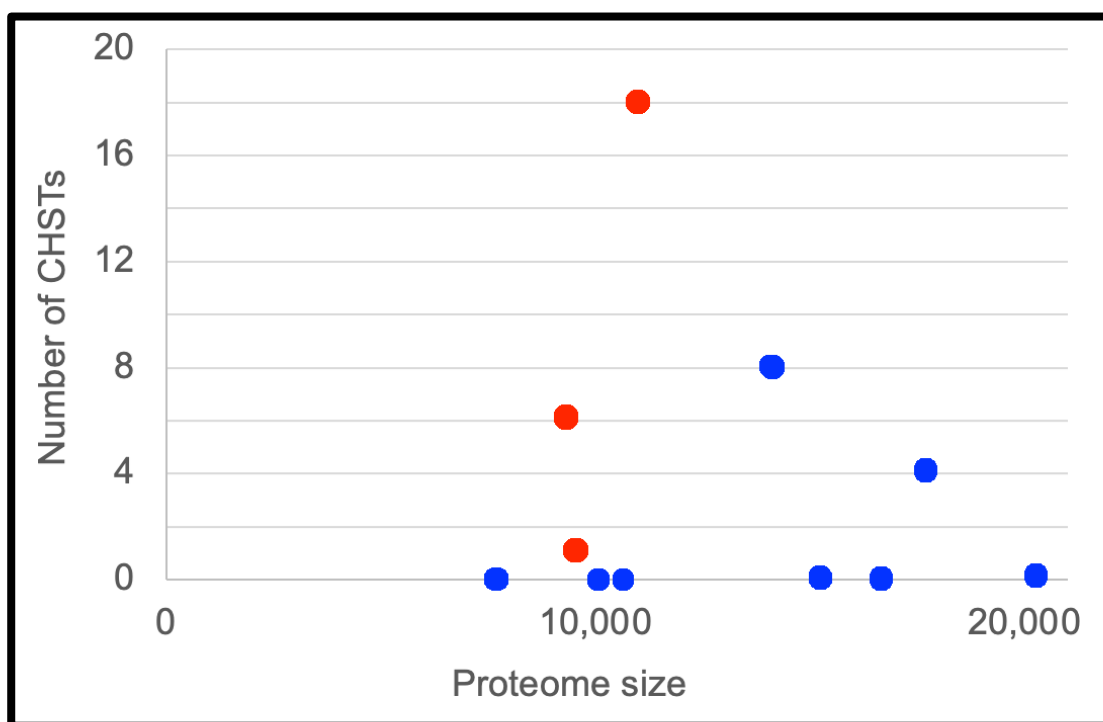


Figure 5.6 Number of carbohydrate SOT as a function of proteome size of select green and red algae. Macroalgae are in red, microalgae are in blue.

After confirming the presence of SOT genes in green algae, an analysis of orthologous genes among diverse green algae was designed. Eight diverse green algal genomes (*B. prasinos*, *C. lentillifera*, *C. reinhardtii*, *Coccomyxa subellipsoidea*, *C. variabilis*, *O. tauri*, *U. compressa*, and *V. carteri*) and one charophyte outgroup species (*C. brauna*) were analyzed via Orthofinder. The program identified a total of 111,839 genes, which parsed into 15,682 orthogroups (Table 5.3). All species have greater than 80% of their genes assigned, except for *U. compressa* (73%). The species range in orthogroup attendance from 32% in the outgroup *C. brauna* to 68% in *C. reinhardtii*.

Table 5.3 Species involved in the Orthofinder analysis and their respective contributions to the total genes assigned and orthogroup count.

Species	Genes assigned (%)	Orthogroups assigned (%)
<i>Bathycoccus prasinos</i>	6,811 (86%)	5,521 (35%)
<i>Chara brauna</i>	31,576 (89%)	5,997 (32%)
<i>Caulerpa lentillifera</i>	8,970 (92%)	5,055 (32%)
<i>Chlamydomonas reinhardtii</i>	17,426 (89%)	10,627 (68%)
<i>Coccomyxa subellipsoidea</i>	8,052 (84%)	6,079 (38%)
<i>Chlorella variabilis</i>	8,771 (90%)	6,079 (39%)
<i>Ostreococcus tauri</i>	6,687 (87%)	5,621 (36%)
<i>Ulva compressa</i>	9,482 (73%)	6,009 (38%)
<i>Volvox carteri</i>	14,064 (88%)	9,864 (63%)
Totals:	111,839 genes into 15,682 orthogroups	

The SOT orthogroups were used to inform a species phylogeny (Figure 5.7) and a phylogeny of the nine aforementioned species and those closely related in the database (Figure 5.8). In the species phylogeny, we see predicted resolution of our current

understanding of species relationships, and accordingly, the charophyte outgroup resolves as such.

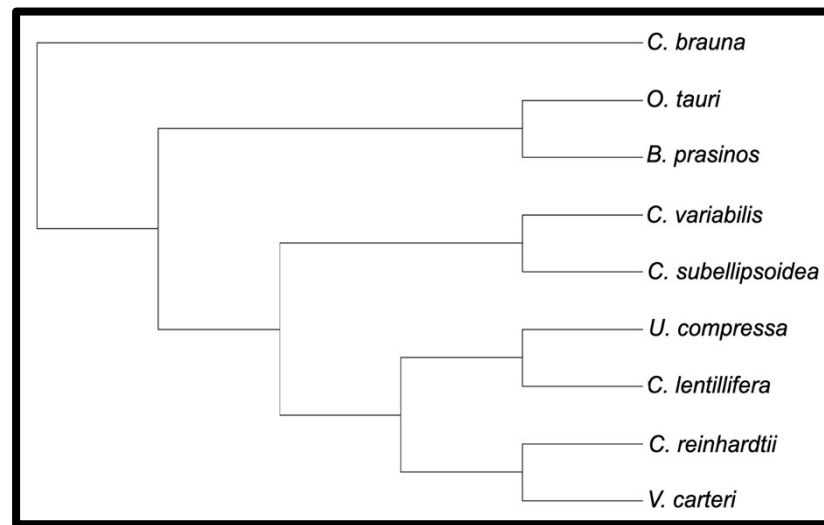


Figure 5.7 Species phylogeny resulting from the orthogroup SOT analysis.

There seem to be two major clades of green CHST-like SOTs. The first is found in soil and freshwater microalgae (*C. reinhardtii*, *V. carteri*) and the second is found in marine macro- and micro-algae (*U. compressa*, *C. lentillifera*, *C. eustigma*). The latter group includes two CHSTs from *Chlamydomonas* UWO241, which is a psychrophilic strain from Antarctica. The differences between the two clades may therefore reflect either substrate specificity or adaptation to salinity. The Akkermansiaceae bacterial sequence is present within the green algae and consistently sits close to *U. compressa* (Figure 5.8). The sole Chondrus CHST-like SOT sits close to the sequences for marine molluscs (*Crassostrea* and *Pomacea*) and marine oömycetes (*Saprolegnia* and *Aphanomcyes*).

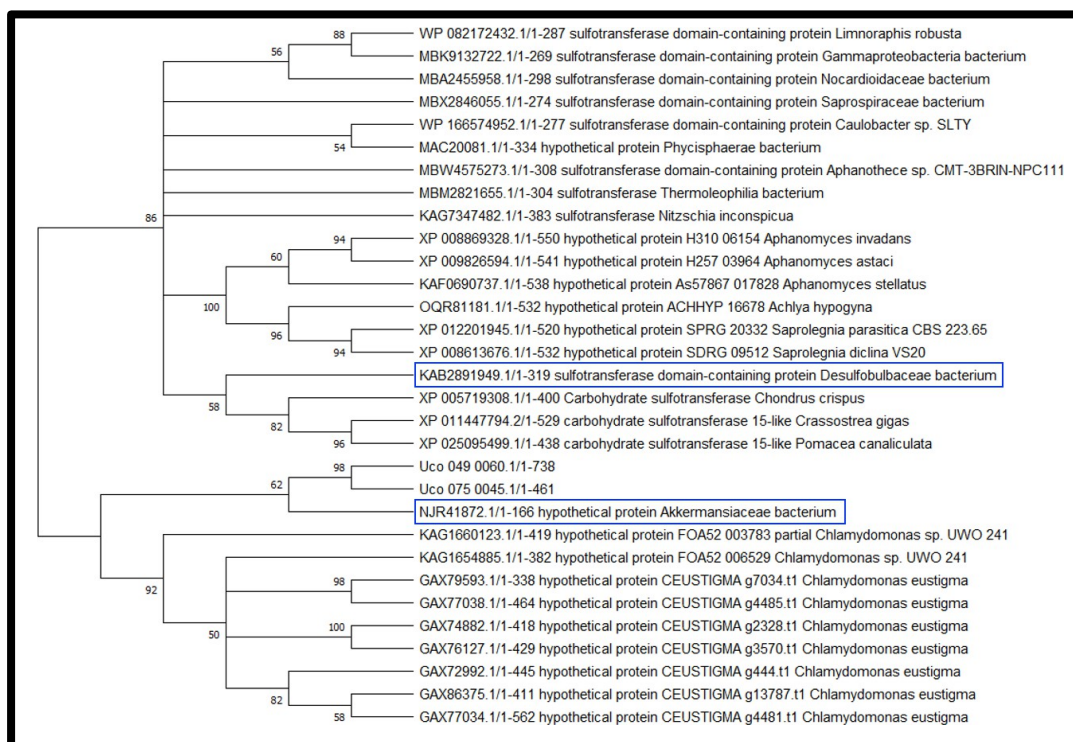


Figure 5.9 Maximum likelihood tree showing rearrangement based on changes of function parameters. Bacterial sequences are highlighted in blue.

With slight changes to analysis parameters, there are rearrangements to clade organization (Figure 5.9). Additionally, these rearrangements alter relationships among algae and bacterial sequences. More specifically, with this exemplar rearrangement, bacteria *Desulfobulbaceae* clusters with *Chondrus*, although this arrangement is notably less stable than the previous example.

5.3.4 Cell wall composition

Based on the obvious segregation of freshwater and marine green algal species with regards to CHST-like sequences, *U. compressa* (marine) and *C. lentillifera* (freshwater) fractions were analyzed via FTIR. The pectin fraction was compared between the two species, as we have previously shown this fraction to contain the most sulfate compounds compared to the remaining extracts (Figure 5.3).

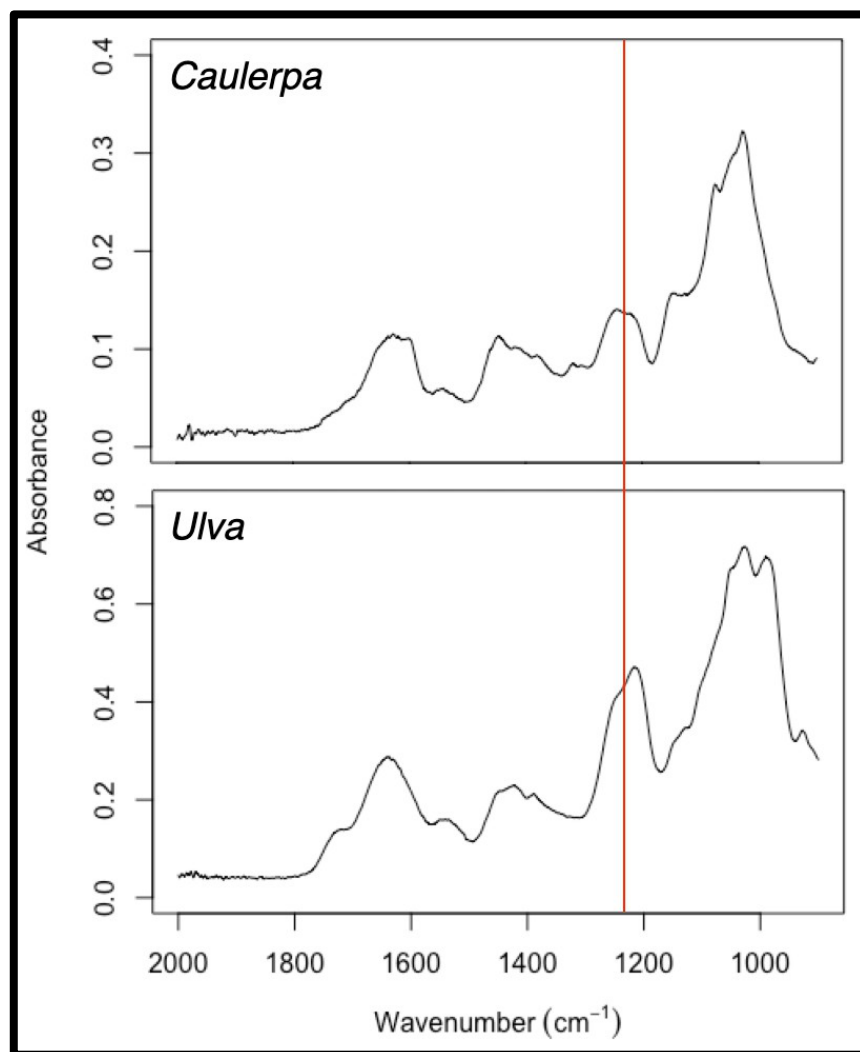


Figure 5.10 FTIR spectroscopy of *Ulva* (upper) and *Caulerpa* (lower). Red line highlights wavenumber $\sim 1230\text{ cm}^{-1}$.

Absorbance of the pectin fraction show nearly 5x sulphate presence in *U. compressa* as compared to *C. lentillifera* based on the height of the peak highlighted in red (Figure 5.10). Although these peaks are complex and have not yet been fully described, there are obvious differences in the elemental analyses with regards to sulphate-containing compounds between these two representative species.

Chapter 6: GENOMIC VARIATION IN ULVACEAE

6.1 Introduction

Ulva is morphologically plastic, making it difficult to visually differentiate one species from another (Bliding 1963, Tanner 1979, Koemann & van den Hoek 1981, Hoeksema & van den Hoek 1983, Blomster et al. 1999, Tan et al. 1999, Hayden & Waaland 2002, Hayden et al. 2003, Hayden & Waaland 2004). Morphology has been shown to change in response to varying environmental conditions (Titlyanov et al. 1975, Tanner 1979, 1986, Malta et al. 1999, Rybak 2018). For example, age, wave exposure, air exposure, temperature, season, salinity, latitude, and whether a frond was sessile or free-floating are associated with morphological variation in thallus thickness, size, and presence/number of marginal teeth.

For this reason, DNA-based methods are the accepted standard for species identification. *Ulva* species have been identified using the chloroplast gene *rbcL* or nuclear ribosomal internal transcribed spacer DNA (ITS2 nrDNA) in a DNA barcoding approach (Saunders and Kucera 2010). While the *Ulva mutabilis* genome has been sequenced, annotation is incomplete. In conjunction with the difficulty in distinguishing species, the use of alternative and less site-specific sequencing methods, such as genotyping-by-sequencing (GBS), is unexplored in *Ulva*. These techniques allow for abbreviated sequencing, a more time and cost-effective strategy.

Given that seaweeds lack seeds and are relatively intensive to maintain in culture, many samples worldwide are stored in collections in desiccated or frozen-form. It is these silica-desiccated or frozen samples onto which these DNA-based methodologies are eventually applied. With the advent of next generation sequencing, and with the ongoing reference whole genome sequencing effort, we determined it prudent to look at whether these stored specimens could also be used to recover and analyze meaningful restriction site-associated DNA sequencing (RADseq) data.

Considering the high degree of genomic variability we know to be present in the genus *Ulva*, an estimation of diversity on a global scale is a novel and worthy focus of study. The goal of the RAD-seq/GBS project was to collect sufficient data of reduced genomic complexity relative to WGS efforts to compare individual *Ulva* samples across various spatial scales. By extracting DNA from the preserved samples and digesting with a cocktail of restriction enzymes, the expected yield would be diverse, yet comparable short read sequences. An overlaying of divergence with the physical characteristics provided with each sample upon arrival in our collection would be a powerful analysis. Not only would this data speak to global *Ulva* diversity and distribution, but it would also highlight small scale selective pressure, and how that pressure manifests in the *Ulva* genome.

6.2 Methods

6.2.1 *Global sample collection and curation*

An initial aim of the project was to create a centralized collection of samples of the genus *Ulva*, which would span as many latitudes and geographical regions as possible. A crude batch of dried samples was transported to Durham University for sorting, curation, and quality evaluation. This batch of samples represents a massive collaboration of biologists, academics, herbariums, and museums worldwide.

Upon receipt of the collection, all samples were logged and weighed. Any other data which accompanied a given sample was filed into a database. The following information was sought for each sample, if available (Table 6.1).

Table 6.1 Data fields sought for each sample in the crude collection.

Concrete data	Proposed data
Name of collector	Phenotype
Country of collection	Species
Sampling site longitude	Site description
Sampling site latitude	Taxa
Date of collection	Haplotype
Storage conditions	<i>rbcL</i> fasta file name and coverage (if applicable)

Samples were sorted based on their relative likeness and were included in the final collection if they met a given set of standards, as outlined in Table 6.2.

Table 6.2 Standards and criteria to be met for inclusion in sample database.

Sample standard	Criterion
Origin	Data must include location of sample collection. (preferably latitude and longitude, but at least country and general region)
Storage/cleanliness	Sample must be dried and stored appropriately with desiccant. Sample must be free of visible epiphytic material.
Integrity	Sample must not show sign of decay or degradation. (by sight and/or smell)
Replicates	Replicate samples will be stored, but not included in the sample database. (replicates are defined as unique individuals from the exact same sampling location)
Mass	Individual samples must be of mass 100mg or greater.
Recovery	Extraction must yield 50 ng DNA with additional tissue remaining for re-extraction.

From these standards, samples were separated into categories, which would inform the proceeding DNA extractions (Table 6.3). These categories were outlined based on the characters of size, quality, and data which accompanied the sample upon receipt.

Table 6.3 Categorical organization of sample types.

Category	Description
A	Of sufficient size and quality
B	Of sufficient size, poor quality
C	Too small, high quality
D	Replicate samples
E	Too little geographic/descriptive data accompanying sample

Samples were then further divided into groups according to location and expected species and were given a unique identification code consisting of one or two letters and one or two numbers. Expected species were provided by the collector of the tissue based on morphological characters, species ranges, and experiences of the collectors themselves.

6.2.2 DNA extractions

All extractions were performed using the DNeasy plant mini kit (Qiagen). This kit is optimized for fresh or frozen samples, and therefore, modifications were necessary to extract DNA from the aged, dried samples of the collection. Specimens for which there was a surplus of tissue provided by the sender were used for trial extractions to adjust the protocol for dried *Ulva* thallus. Each iteration was performed from one specimen for consistency. The protocol steps of tissue homogenization, starting sample mass, and volumes of reagents were modified (Table 6.4).

Table 6.4 DNeasy protocol modifications for tissue-specific extractions.

Protocol step	Modification	Iteration
Tissue homogenization	Motorized tissue ruptor with disposable/sterile pellet pestle	Dry or immersed in liquid nitrogen
	Thistle bullet blender with beads and lysis buffer	Nuclease-free water or lysis buffer
Initial mass	Thallus mass	20, 30, 35 or 40 mg
	Lysis buffer volume (Ap1)	400 500, 600, 700, or 800 μ L (proportional to mass of 1, 1.5, 1.75, or 2x)
	RNase volume (A)	4, 6,7, or 8 μ L (proportional to mass of 1, 1.5, 1.75, or 2x)
	Neutralization buffer volume (P3)	130, 195, 228, 260 μ L (proportional to mass of 1, 1.5, 1.75, or 2x)
Elution	Replicates	1 or 2
	Incubation time	5, 10, or 15 minutes
	Incubation volume	50 or 100 μ L
	Elution matrix	Buffer AE or nuclease-free water

The target DNA concentration was at least 50 ng/ μ L and the target purity was an A260/A280 greater than 1.4. Concentration and purity were tested in duplicate using Nanodrop. Changes to the protocol were individually and systematically tested, and these parameters were compared to decide on a final version. Samples were stored at -20°C after testing.

6.2.3 Restriction site-associated DNA sequencing

This collection was the basis of a proposed RAD-seq analysis to understand the biogeography of *Ulva*, especially considering the extreme range of suitable habitats of this genus. A subset of samples which met acceptable standards were identified as an experimental batch to test efficacy of the RADseq protocol with these sample types. Thirty high quality and large mass individuals were transported to Aberystwyth University for the trial run with 90 unique restriction site tags. Unaligned sequences were returned for 20 individuals, which were then aligned to the reference genome, *U. mutabilis*, using Bowtie. The unaligned and aligned sequences were analyzed for identity, agreement between samples, and whether they corresponded to restriction

sites from the intended analysis. Results from both the aligned and unaligned fasta files were inputted to BLAST to annotate the sequences in terms of the reference genome (NCBI).

6.3 Results and Discussion

6.3.1 *Modified extraction protocol*

The final parameters selected for the optimized protocol include changes to initial mass, homogenization method, elution sequence, and lysis buffer volume (Table 6.5). These parameters would be used henceforth in any future extractions for dried *Ulva* thallus.

Table 6.5 Description of protocol modifications for dried sample extractions and corresponding test results.

Parameter	Measure/Technique	
	DNeasy Plant mini kit	Optimized protocol
Sample mass	20 mg	30 mg
Homogenization method	Motorized tissue ruptor with disposable/sterile pellet pestle	Thistle bullet blender with beads and lysis buffer
Elution sequence	2 x 100 μ L (5 min. incubation)	2 x 50 μ L (10 min. incubation)
Buffer AP1 (lysis buffer)	400 mL	500 mL
Results	DNA concentration	4.7-9.6 ng/ μ L
	260/280	1.3
		45.1-49.3 ng/ μ L
		1.7

Initial thallus used in each extraction was increased 50% from 20 mg to 30 mg. This tissue was homogenized using a Thistle Scientific brand bullet blender with metallic beads. This process essentially combined the mechanical homogenization of the pellet pestle with the biochemical homogenization of buffer AP1 by adding lysis buffer directly to the sample tube in the machine. The elution volume was decreased by 50% from 100 μ L to 50 μ L, and incubation time was increased 50% from 5 minutes to 10 minutes. This change in the process resulted in a smaller volume of more concentrated filtrate (Table 6.5).

These parameters resulted in a 5-10 fold increase in estimated DNA concentration recovery and an in-range A260/A280 reading (Figure 6.5). With these changes, the goal parameters for DNA extractions were consistently achieved over different thallus samples.

6.3.2 *Ulva* Genotyping-by-Sequencing

From an original collection of over 500 specimens, the database of usable, extract-able thalli was refined to 273 individuals (Table 6.6). Samples filtered out from the final database were too small, of unacceptable quality, both too small and of unacceptable quality, and those which were not accompanied by any geographic descriptive data. Replicate samples were not removed, but rather were parsed into a separate collection as a secondary biomass resource.

Table 6.6 Presentation of unique specimens in the global collection in order of ascending northern latitude. DNP = data not provided by source.

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
LA24-5	New Zealand	-45.949	DNP	May 2005	S. Heesch
LA27	New Zealand	-45.875	DNP	May 2005	S. Heesch
LA28	New Zealand	-45.853	DNP	May 2005	S. Heesch
LA26	New Zealand	-45.813	DNP	May 2005	S. Heesch
CN8	Chile	-42.469	DNP	Jan 2009	J. Brodie
CN3	Chile	-41.902	DNP	Jan 2009	J. Brodie
CN6	Chile	-41.902	DNP	Jan 2009	J. Brodie
CO7	Chile	-41.902	DNP	Jan 2009	J. Brodie
LA17	Chile	-41.505	DNP	DNP	T. Alsufyani
NC5-7	Chile	-41.492	DNP	Jan 2009	J. Brodie
LA16	Chile	-41.467	DNP	Oct 2000	H.S. Hayden
LA21-2	New Zealand	-41.338	DNP	Mar 2005	S. Heesch
LA23	New Zealand	-41.298	DNP	Mar 2005	S. Heesch
LA18-9	Australia	-38.293	<i>stenophylloides</i>	Sep 2006	L.G.K. raft
LA20	Australia	-38.270	DNP	Dec 2006	L.G.K. raft
DX1	South Africa	-34.832	DNP	Nov 2010	V. Cole
DW7	South Africa	-34.177	DNP	Dec 2010	V. Cole
NA9	South Africa	-34.164	DNP	Apr 2015	O. De Clerck
DW9	South Africa	-34.113	DNP	Nov 2010	V. Cole
DW5	South Africa	-34.054	DNP	Dec 2010	V. Cole
EC8	South Africa	-33.902	<i>capensis</i>	Jan 1851	J.E. Areschoug
DX6	South Africa	-31.986	DNP	Nov 2010	V. Cole
CF3	Australia	-31.833	DNP	Mar 2008	C. McInerney
CF8	Australia	-31.833	DNP	Mar 2008	C. McInerney
DW1	South Africa	-31.744	DNP	Nov 2010	V. Cole
DW3	South Africa	-31.744	DNP	Nov 2010	V. Cole
BX2	South Africa	-30.986	<i>fasciata</i>	Aug 1999	DNP
BX4	South Africa	-30.250	DNP	Aug 1999	DNP
DC0	Australia	-30.109	DNP	Nov 2009	H. Verbruggen
DR7-8	Brazil	-23.531	DNP	Sep 2010	H. Verbruggen
DR6	Brazil	-23.507	DNP	Sep 2010	H. Verbruggen
BW6	Madagascar	-23.401	<i>reticulata</i>	Aug 2002	DNP
BY3	Madagascar	-23.356	<i>lactuca</i>	Aug 2002	DNP
CL4	Brazil	-23.068	<i>lactuca</i>	Feb 2008	B. Castelar
CL6	Brazil	-23.068	<i>lactuca</i>	Feb 2008	B. Castelar
CL8	Brazil	-23.068	DNP	Feb 2008	B. Castelar
CL5	Brazil	-22.964	<i>fasciata</i>	Mar 2008	B. Castelar
CL7	Brazil	-22.964	<i>fasciata</i>	Mar 2008	B. Castelar
CE2	Namibia	-22.812	DNP	Feb 2008	L. Kandjengo
DR0	Brazil	-22.762	DNP	Sep 2010	H. Verbruggen
DR2	Brazil	-22.762	DNP	Sep 2010	H. Verbruggen
DR4-5	Brazil	-22.762	DNP	Sep 2010	H. Verbruggen
CE7	Namibia	-22.698	<i>nematoidea</i>	Feb 2008	L. Kandjengo

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
BP6	New Caledonia	-22.675	DNP	Jun 2007	S. Andrefouet
BO1	New Caledonia	-22.671	DNP	Nov 2005	L. Mattio
CD8	Namibia	-22.623	<i>nematoidea</i>	Feb 2008	L. Kandjengo
BN9	New Caledonia	-22.584	DNP	Dec 2001	C. Payri
BU6-7	New Caledonia	-22.575	DNP	Sep 2007	C. Payri
BO5	New Caledonia	-21.049	DNP	Mar 2005	C. Geoffray
DT2	Reunion Island	-21.026	DNP	DNP	M. Zubia
DT4	Reunion Island	-21.026	DNP	DNP	M. Zubia
DT6	Reunion Island	-21.026	DNP	DNP	M. Zubia
EE1	Reunion Island	-21.026	DNP	DNP	M. Zubia
FM12-3	Reunion Island	-21.026	DNP	DNP	M. Zubia
BX1	Reunion Island	-19.782	DNP	Oct 2001	DNP
AZ9	Peru	-18.296	DNP	Jun 1905	C.Maggs
CL3	Moorea	-17.484	<i>ranunculata</i>	Oct 2008	A. N'Yeurt
BZ0	Moorea	-17.481	DNP	DNP	DNP
ED3	Madagascar	-14.937	DNP	May 2012	L. Mattio
CL9	Brazil	-12.939	DNP	Feb 2008	B. Castelar
NB9	Tanzania	-10.305	DNP	Jan 2008	H. Verbruggen
NC4	Tanzania	-10.305	<i>reticulata</i>	Jan 2008	K. Pauly
NC0	Tanzania	-10.305	DNP	Jan 2008	H. Verbruggen
CI2	Tanzania	-10.288	<i>reticulata</i>	Jan 2008	H. Verbruggen
BW8	Tanzania	-10.267	<i>reticulata</i>	Jul 2000	DNP
DK4	Indonesia	-8.910	DNP	Jun 2010	DNP
CU8	Indonesia	-8.895	<i>lactuca</i>	Apr 2006	DNP
CW0	Indonesia	-8.895	DNP	Apr 2006	DNP
CU4	Indonesia	-8.620	DNP	Aug 2005	DNP
DK3	Indonesia	-8.468	DNP	Jun 2010	DNP
DA0	Indonesia	-8.111	DNP	Jan 2010	I.S. Arlyza
DA2	Indonesia	-8.111	DNP	Jan 2010	I.S. Arlyza
DP6	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
DP7	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
DP8	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
DQ0	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
DQ2	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
DQ3-7	Ascension Island	-7.987	DNP	Nov 2010	O. Renshaw
NC3	Tanzania	-6.666	DNP	Jan 2008	H. Verbruggen
NC2	Tanzania	-6.666	DNP	Jan 2008	H. Verbruggen
NC1	Tanzania	-6.662	DNP	Jan 2008	H. Verbruggen
CI1	Tanzania	-6.659	<i>reticulata</i>	Jan 2008	K. Pauly
DD1-2	Tanzania	-6.414	DNP	Feb 2008	H. Verbruggen
DD4	Tanzania	-6.414	DNP	Feb 2008	H. Verbruggen
DD5	Tanzania	-6.268	DNP	Feb 2008	H. Verbruggen
BL1-9	Tanzania	-6.165	<i>reticulata</i>	Jul 1997	DNP
DE1	Tanzania	-6.153	DNP	Feb 2008	H. Verbruggen
NB6	Tanzania	-6.153	DNP	Feb 2008	H. Verbruggen
DD7	Tanzania	-5.877	<i>lactuca</i>	Feb 2008	H. Verbruggen

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
NB5	Tanzania	-5.829	DNP	Feb 2008	H. Verbruggen
NB7-8	Tanzania	-5.725	DNP	Feb 2008	H. Verbruggen
BR6	Kenya	-4.663	<i>reticulata</i>	Mar 1988	E. Coppejans
CG9	Kenya	-4.287	<i>reticulata</i>	Mar 2008	O. De Clerck
NA4	Kenya	-3.918	DNP	Oct 2011	L. Mattio
BS8	Kenya	-3.821	<i>reticulata</i>	Jul 1989	E. Coppejans
DG5	Indonesia	0.779	DNP	Sep 2009	S. Draisma
DA4	Indonesia	0.853	DNP	Nov 2009	I.S. Arlyza
DF6	Indonesia	1.252	DNP	DNP	DNP
DF7	Indonesia	1.252	DNP	DNP	DNP
DG3	Indonesia	1.252	DNP	DNP	DNP
CX1	Malaysia	1.316	DNP	Aug 2009	L.P. Eem
CX3	Malaysia	1.316	<i>umbraulva</i>	Aug 2009	L.P. Eem
DK2	Malaysia	1.436	DNP	May 2010	DNP
DH4	Malaysia	2.117	<i>fasciata</i>	Nov 2009	F.Mineur
DH2	Malaysia	2.435	<i>reticulata</i>	Apr 2010	DNP
NA0	Malaysia	4.210	DNP	Apr 2010	DNP
BW4	Sri-Lanka	6.009	<i>laciniata</i>	Jan 2000	DNP
BY1	Sri-Lanka	6.009	DNP	Jan 2000	DNP
BW1	Sri-Lanka	6.129	<i>reticulata</i>	Jan 1997	DNP
BX6	Sri-Lanka	6.187	<i>fasciata</i>	Feb 2007	DNP
BX7	Sri-Lanka	6.187	DNP	Feb 2007	DNP
DT9	Benin	6.357	DNP	Dec 2010	M. Uchimura
DU1	Benin	6.357	DNP	Dec 2010	M. Uchimura
CH2	Tanzania	6.369	<i>reticulata</i>	DNP	DNP
BT1	Philippines	6.905	<i>reticulata</i>	Jan 1876	J. Moseley
BW9	Sri-Lanka	7.560	DNP	Sep 2007	DNP
BT9	Philippines	9.224	<i>reticulata</i>	Sep 2007	O. De Clerck
BU4-7	Philippines	9.224	<i>reticulata</i>	Sep 2007	O. De Clerck
BU1	Philippines	9.331	DNP	Sep 2007	O. De Clerck
LA3	Philippines	10.285	<i>reticulata</i>	Apr 1998	S. Shimada
BZ9	Philippines	10.302	<i>reticulata</i>	Jan 2004	DNP
UC1-10	Curacao	12.170	<i>curacao</i>	Jun 2012	DNP
BD1	Yemen	13.320	<i>reticulata</i>	Nov 1904	P. Forsskal
CZ0	Senegal	14.179	DNP	Dec 2009	M. Uchimura
CZ2	Senegal	14.179	DNP	Dec 2009	M. Uchimura
CZ4	Senegal	14.285	DNP	Dec 2009	M. Uchimura
CZ6	Senegal	14.285	DNP	Dec 2009	M. Uchimura
CY6	Senegal	14.738	DNP	Dec 2009	M. Uchimura
CY8	Senegal	14.738	DNP	Dec 2009	M. Uchimura
DJ9	West Indies	16.228	<i>laetevirens</i>	Jan 1890	British Museum
LA9	India	17.439	<i>ohnoi</i>	Apr 2014	M.A. Kazi
LA12	India	17.628	<i>ohnoi</i>	Sep 2014	M.A. Kazi
LA8	India	17.644	<i>ohnoi</i>	Apr 2013	M.A. Kazi
LA11	India	17.721	<i>ohnoi</i>	Nov 2013	M.A. Kazi
LA13	India	17.989	<i>ohnoi</i>	Dec 2013	M.A. Kazi

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
LA15	India	17.989	<i>ohnoi</i>	Dec 2013	M.A. Kazi
DJ3	USVI (St. Thomas)	18.338	<i>lactuca</i>	DNP	British Museum
BX0	Dominican Republic	19.829	DNP	Feb 2002	DNP
LA7	USA (Hawaii)	20.783	<i>reticulata</i>	Feb 2000	H.S. Hayden
LA10	India	20.902	<i>ohnoi</i>	Dec 2013	M.A. Kazi
ND6-7	New Caledonia	20.904	<i>ohnoi</i>	Oct 2006	DNP
ND8-9	New Caledonia	20.904	<i>ohnoi</i>	Oct 2006	DNP
DK5	USA (Hawaii)	21.277	DNP	Jul 2009	C. Maggs
DK7	USA (Hawaii)	21.277	<i>reticulata</i>	Jul 2009	C. Maggs
DK8	USA (Hawaii)	21.277	DNP	Jul 2009	C. Maggs
DK9	USA (Hawaii)	21.593	DNP	Jul 2009	C. Maggs
LA4	India	22.451	<i>reticulata</i>	DNP	V. Gupta
BY5	Oman	22.638	<i>rigida</i>	Sep 2003	T. Schils
BZ1	Oman	22.638	DNP	Feb 2003	T. Schils
ND3-4	Brazil	22.962	<i>fasciata</i>	Mar 2008	DNP
ND5	Brazil	23.069	<i>lactuca</i>	Feb 2008	DNP
BY6	Oman	23.522	DNP	Sep 2003	T. Schils
LA37	Japan	26.224	<i>ohnoi</i>	Dec 1998	M. Hiraoka
LA6	Japan	26.224	<i>ohnoi</i>	Dec 1998	M. Hiraoka
DB5	Japan	26.329	DNP	Jul 2009	K. Pauly
DB7	Japan	26.329	DNP	Jul 2009	K. Pauly
ED7	Japan	26.385	DNP	DNP	L. Mattio
ED8	Japan	26.385	DNP	DNP	L. Mattio
EE0	Japan	26.385	DNP	DNP	L. Mattio
DB4	Japan	26.442	DNP	Jul 2009	K. Pauly
DB3	Japan	26.444	DNP	Jul 2009	K. Pauly
CC0	USA (Florida)	27.255	DNP	Aug 2004	DNP
DF4	USA (Texas)	27.842	<i>fasciata</i>	DNP	M. Wynne
BE1	Spain	28.418	<i>fasciata</i>	Feb 2003	C. Maggs
BP7	Spain	28.420	<i>rotundata</i>	Mar 1964	C. Bliding
DI5	Spain	28.744	DNP	Mar 2010	F. Mineur
CJ4	Egypt	29.548	DNP	DNP	J.M. Erik
CJ6	Egypt	29.548	DNP	DNP	J.M. Erik
CA0	USA (California)	29.954	DNP	DNP	DNP
BH8	Portugal	32.641	<i>fasciata</i>	Dec 2006	C. Maggs
BH5	Portugal	32.641	<i>lactuca</i>	Dec 2006	C. Maggs
LA5	USA (California)	32.695	<i>taeniata</i>	DNP	H.S. Hayden
LA33	Japan	33.452	<i>ohnoi</i>	Feb 2000	M. Hiraoka
LA35	Japan	33.452	<i>ohnoi</i>	May 1999	M. Hiraoka
LA36	Japan	33.452	<i>ohnoi</i>	Sep 1997	M. Hiraoka
LA2	Japan	33.473	<i>spinulosa</i>	Apr 1999	S. Shimada
LA34	Japan	33.598	<i>ohnoi</i>	Sep 1999	M. Hiraoka
LA1	Japan	35.222	<i>conglobata</i>	Jun 2013	K. Matsumoto
NA3	Malta	36.053	DNP	Jul 2012	C. Maggs
NA1	Malta	36.072	DNP	Jul 2012	DNP

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
BM4	Spain	36.482	DNP	Apr 2007	J. Lucas
LA29	Spain	36.524	<i>clathrata</i>	May 1998	M. Blomster
LA32	Spain	36.524	<i>muscoides</i>	May 1998	M. Blomster
BG0	Spain	36.528	<i>rigida</i>	Nov 2006	F.Mineur
BD3	USA (California)	36.614	<i>fasciata</i>	Jun 1901	Setchell
CN0	USA (California)	36.614	<i>fasciata</i>	Mar 1905	Setchell
NA7	Tunisia	36.717	DNP	Dec 2010	H. Langar
NA6	Tunisia	36.720	DNP	Dec 2010	H. Langar
DY3	Tunisia	36.774	DNP	Dec 2010	H. Langar
NA5	Tunisia	36.822	DNP	Dec 2010	H. Langar
NA4	Tunisia	36.832	DNP	Dec 2010	H. Langar
CB6	Portugal	37.003	DNP	Dec 2007	C. Maggs
AQ2	Portugal	37.006	<i>rotundata</i>	Dec 2005	E.J. Malta
LA14	Portugal	37.020	<i>ohnoi</i>	May 2010	T. Alsufyani
LA30	Portugal	37.020	<i>muscoides</i>	May 2010	T. Alsufyani
LA31	Portugal	37.020	<i>muscoides</i>	May 2010	T. Alsufyani
AQ4	Portugal	37.092	<i>rotundata</i>	Jun 2005	E.J. Malta
DX9	Greece	37.389	DNP	Sep 2011	F. Mineur
DY1	Greece	37.446	DNP	Sep 2011	F. Mineur
CK2	Italy	38.238	DNP	May 2008	DNP
NB0-2	Italy	38.263	DNP	May 2008	DNP
CK4	Italy	38.264	DNP	May 2008	DNP
CK6	Italy	38.264	DNP	May 2008	DNP
NC8-9	Italy	38.2646	DNP	DNP	DNP
ND0-1	Italy	38.2646	DNP	DNP	DNP
CF1	Italy	40.282	<i>rigida</i>	Mar 2008	A. Soro
AY4	Italy	40.829	<i>rotundata</i>	Apr 1958	C. Bliding
CA3	Spain	41.862	DNP	Jul 2006	DNP
DM2	France	42.483	DNP	Jul 2010	F.Mineur
CI6	France	42.566	<i>rigida</i>	Jun 2008	V. Demoulin
CI8	France	42.566	<i>rigida</i>	Jun 2008	V. Demoulin
CJ0	France	42.580	<i>rigida</i>	Jun 2008	V. Demoulin
BJ6	France	43.241	<i>rigida</i>	Mar 2007	M.Verlaque
AZ5	France	43.430	<i>fasciata</i>	Sep 2002	M. Verlaque
BM2	France	43.496	<i>rigida</i>	May 2007	B. Guillaume
BK8	France	43.512	<i>rigida</i>	May 2007	B. Guillaume
NA8	Spain	43.529	DNP	Jul 2008	C. Maggs
CH7	USA (Oregon)	44.497	<i>taeniata</i>	May 2008	G. Hansen
NB3-4	France	44.6523	DNP	DNP	I. Fremer
CE8	France	44.666	DNP	DNP	I. fremer
DL4	Italy	45.420	DNP	May 2010	M. Verlaque
DL8	Italy	45.420	DNP	May 2010	M. Verlaque
DL6	Italy	45.420	DNP	May 2010	M. Verlaque
BF7	Slovenia	45.540	<i>laetevirens</i>	DNP	DNP
BF5	Slovenia	45.540	<i>scandinavica</i>	DNP	DNP
BD5	France	45.982	<i>fasciata</i>	Jul 2006	F. Mineur

Unique ID	Collection Location	Northern Latitude	Proposed species	Date of collection	Collector or provider
NA2	France	46.048	<i>rigida</i>	Aug 2010	F. Mineur
ND2	France	48.7262	<i>rigida</i>	Nov 2007	C. Maggs
BS9	France	48.726	DNP	Sep 2007	M. Valero
DN2	France	48.729	DNP	Jul 2010	J. Bothwell
DN0	France	48.729	DNP	Jul 2010	J. Bothwell
DL2	England	50.830	DNP	Jul 2010	C. Maggs
CA4	Wales	51.775	DNP	Dec 2007	C. Maggs
DN4	Wales	51.870	DNP	Aug 2010	C. Maggs
DN6	Wales	51.870	DNP	Aug 2010	C. Maggs
DN8	Wales	51.870	DNP	Aug 2010	C. Maggs
DO0	Wales	51.870	DNP	Aug 2010	C. Maggs
DO2	Wales	51.870	DNP	Aug 2010	C. Maggs
DO4	Wales	51.870	DNP	Aug 2010	C. Maggs
DL3	Northern Ireland	54.373	DNP	Jul 2010	C. Maggs
BM8	Ireland	54.373	DNP	Jun 2007	C. Maggs
BN2	Ireland	54.373	DNP	Jun 2007	C. Maggs
BM7	Northern Ireland	54.373	DNP	Jun 2007	C. Maggs
BJ5	Ireland	55.041	<i>pseudocurvata</i>	Mar 2007	C. Maggs
BH3	Ireland	55.201	<i>clathrata</i>	Dec 2006	C. Maggs
BQ0	Norway	60.248	<i>scandinavica</i>	Oct 1962	C. Bliding
BP9	Norway	60.321	<i>scandinavica</i>	Oct 1962	C. Bliding

Samples from Table 6.6 represent 47 distinct countries (or regions, in few cases), each with between 1-19 unique samples originating from various coastal environments within that region. A global view of these regions (Figure 6.1) provides a comprehensive visual of that distribution, both in the origin of samples in the collection and of the many homes made around the world by *Ulva* spp.



Figure 6.1 Mapped locations of samples compiled in the global collection. Red dots represent unique countries for which samples are accounted in the collection (most dots represent an average of 5 unique coastal specimens). Bar at lower right equals 500 km (Google).

Amassing this expansive collection was truly a feat of global collaboration. Representing the largest known assortment of *Ulva* specimens, this database serves as both an indication of the diversity of the genus and the power of scientific partnership. Samples in the collection range in latitude from -45 to 60°N, covering an astounding 80% of non-Arctic or Antarctic latitudes. With data on exact GPS coordinates from the providers, we can also confirm the specimens cover a massive expanse of ecological backdrops; from temperate rocky shores to tropical sandy beaches, and metropolitan-adjacent green tides. This quality of the collection lends itself to a broad biogeographic analysis, such as RADseq, sometimes referred to more broadly as GBS, or genotyping-by-sequencing.

A 30-sample subset of this data was sent for a trial analysis to collaborators in Aberystwyth. Samples in the collection were randomly selected from the subset which contained sufficient tissue of acceptable quality and multiple individual specimens. Returned was a fasta alignment of a candidate sequence from 20 samples which aligned to the reference genome (Figure 6.2), a fasta alignment of a candidate sequence from 30 samples which did not align to the reference genome, and details of reference

genome hits from the aligned sequences.



Figure 6.2 Mapped locations of samples in RADseq analysis. Red dots represent exact locations of sample origin. Bar at lower right equals 500 km (Google).

Of the 20 aligned candidates, 18 provided protein (BLASTX) or coding sequence (BLASTN) (or both) hits to the reference genome, *U. mutabilis* (Table 6.7). These sequences provided 97 unique hits to genes on the reference genome spanning representatives on 71 contigs (Table 6.7). Four genes, UM0061_0009.1, UM0061_0010.1, UM0033_0060.1, and UM0120_0012.1 were hit multiple times by different samples and code for a gene involved in RNA methylation, a gene involved in protein binding, a membrane-associated protein, and a Class I transposable element, respectively. Two of the 20 aligned to the reference genome but provided no hits to a gene. These samples are thought to have recovered some kind of non-coding DNA, although this is an unexpected result based on the selected restriction tags and the 90% hit rate for coding sequence or protein data from the remaining 18 candidates.

Table 6.7 Database identifications of sequences from aligned RADseq fasta files.
Samples were randomly selected from a pool of high-quality candidates (Table 6.5).

Random Sample ID	Gene ID	Bit score / E-value	Database
1	UM109_0039.1	44 / 1E-5	Protein / CDS
	UM0061_0009.1	43 / 2E-05	Protein / CDS
2	UM0061_0010.1	33 / 0.017	Protein / CDS
	UM0033_0060.1	37 / 0.005	CDS
3	UM0061_0009.1	43 / 2E-05	Protein / CDS
	UM0061_0010.1	33 / 0.017	Protein / CDS
	UM0033_0060.1	37 / 0.005	CDS
	UM0001_0212.1	50 / 1E-07	Protein / CDS
	UM0034_0020.1	32 / 0.029	CDS
	UM0064_0030.1	32 / 0.038	CDS
	UM0027_0049.1	30 / 0.11	Protein / CDS
4	UM0012_0074.1	28 / 0.55	CDS
	UM0005_0183.1	28 / 0.55	CDS
	UM0239_0006.1	28 / 0.71	CDS
	UM0184_0006.1	27 / 0.93	CDS
	UM0001_0111.1	27 / 0.93	CDS
	UM0001_0265.1	104 / 4E-23	CDS
	UM0077_0045.1	48 / 3E-6	CDS
5	UM0204_0006.1	37 / 0.005	CDS
	UM0208_0008.1	36 / 0.019	CDS
	UM0126_0020.1	36 / 0/019	CDS
	UM0029_0100.1	36 / 0/019	CDS
	UM0027_0035.1	36 / 0/019	CDS
	UM0027_0005.1	36 / 0/019	CDS
	UM0036_0032.1	39 / 0.002	CDS
	UM0062_0035.1	37 / 0.005	CDS
6	UM0044_0006.1	37 / 0.005	CDS
	UM0063_0029.1	36 / 0.019	CDS
	UM0009_0090.1	36 / 0.019	CDS
7	UM0009_0098.1	54 / 1E-8	Protein / CDS
	UM0002_0005.1	36 / 0.019	CDS
8	UM0025_0008.1	29 / 0.24	Protein
	UM0030_0012.1	36 / 0.019	CDS
9	UM0024_0025.1	51 / 8E-8	Protein / CDS
	UM0222_0001.1	36 / 1.9E-2	CDS
	UM0005_0148.1	59 / 4E-10	Protein / CDS
	UM0017_0106.1	37 / 0.002	Protein
	UM0001_0202.1	33 / 0.022	Protein / CDS
	UM0145_0017.1	32 / 0.029	Protein / CDS
10	UM0103_0032.1	32 / 0.049	Protein
	UM0019_0091.1	32 / 0.049	Protein / CDS
	UM0011_0101.1	31 / 0.064	Protein / CDS

Random Sample ID	Gene ID	Bit score / E-value	Database
	UM0004_0056.1	31 / 0.084	Protein
	UM0095_0011.1	30 / 0.11	Protein
	UM0014_0121.1	30 / 0.11	Protein / CDS
	UM0241_0002.1	30 / 0.14	Protein
	UM0021_0013.1	30 / 0.14	Protein
	UM0002_0087.1	30 / 0.14	Protein
	UM0218_0032.1	30 / 0.19	Protein
	UM0031_0016.1	30 / 0.19	Protein
	UM0348_0005.1	29 / 0.24	Protein
	UM0161_0021.1	29 / 0.24	Protein
	UM0017_0062.1	29 / 0.24	Protein
	UM0183_0013.1	29 / 0.24	Protein
	UM0013_0076.1	29 / 0.32	Protein
	UM0010_0133.1	29 / 0.32	Protein
	UM0006_0044.1	29 / 0.32	Protein
	UM0249_0002.1	29 / 0.32	Protein
	UM0139_0005.1	29 / 0.42	Protein
	UM0002_0191.1	29 / 0.42	Protein
	UM0120_0015.1	29 / 0.42	Protein
	UM0025_0040.1	28 / 0.55	Protein
	UM0235_0001.1	28 / 0.55	Protein
	UM0170_0002.1	28 / 0.71	Protein
	UM0073_0011.1	28 / 0.71	Protein
	UM0049_0023.1	28 / 0.71	Protein
	UM0045_0036.1	28 / 0.71	Protein
	UM0119_0047.1	28 / 0.71	Protein
	UM0072_0023.1	27 / 0.93	Protein
	UM0044_0032.1	27 / 0.93	Protein
	UM0002_0131.1	27 / 0.93	Protein
	UM0001_0039.1	27 / 0.93	Protein
	UM0052_0045.1	50 / 9E-7	CDS
	UM0013_0122.1	43 / 1E-4	CDS
	UM0010_0115.1	43 / 1E-4	CDS
	UM0144_0025.1	37 / 0.005	CDS
	UM0358_0001.1	36 / 0.019	CDS
	UM0032_0024.1	36 / 0.019	CDS
11	UM0101_0015.1	43 / 1E-4	CDS
	UM0052_0066.1	36 / 0.018	CDS
12	UM0117_0011.1	36 / 0.019	CDS
15	UM0120_0012.1	57 / 1E-9	Protein / CDS
16	UM0120_0012.1	57 / 1E-9	Protein / CDS
	UM0046_0001.1	39 / 0.002	CDS
	UM0014_0103.1	39 / 0.002	CDS
17	UM0007_0067.1	39 / 0.002	CDS
	UM0015_0028.1	37 / 0.005	CDS
	UM0080_0018.1	36 / 0.019	CDS

Random Sample ID	Gene ID	Bit score / E-value	Database
18	UM0025_0035.1	36 / 0.019	CDS
	UM0075_0013.1	68 / 9E-13	CDS
	UM0039_0069.1	104 / 4E-23	CDS
19	UM0082_0011.1	39 / 0.001	CDS
	UM0207_0011.1	36 / 0.015	CDS
	UM0057_0005.1	36 / 0.015	CDS
	UM0008_0169.1	36 / 0.015	CDS
	UM0008_0131.1	31 / 0.084	Protein / CDS
20	UM0020_0033.1	48 / 3E-6	CDS

Of all aligned hits to coding sequence and protein databases, bit score and e-value range from 27 and 0.93 on the low end and 104 and 4.0E-23 on the high end, respectfully. As the quality scores in this dataset demonstrate a spectrum in usability, those hits with E-value less than or equal to 0.05 were extracted (Figure 6.3). This subset comprised 63% of the original data.

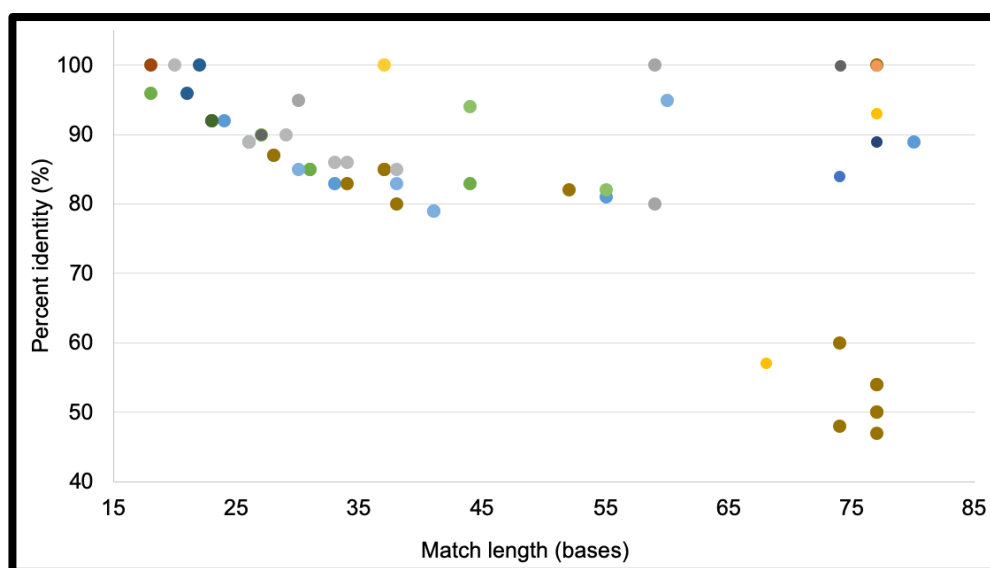


Figure 6.3 Match length and percent identity of gene hits which returned significant E-values. Colors represent the random sample ID.

Significant gene hits to the reference genome seem to group bi-modally. One subset shows very high identity (80-100%) with very short match lengths (~20-55 bases). The other subset shows relatively longer match lengths (~70-80 bases) and varying percent identity (50-100%).

The files that returned unaligned provided far fewer gene hits. Protein database genes UM_0006_128.1, UM0057_0034.1, and UM0056_0028.1 were found in samples seven, 12, and 13, respectively. Sample 13 was one which did not match any gene hits to the reference genome. No coding sequence alignments were identified in this subset. Intuitively, those that do not align should not provide hits.

Within the entire GBS dataset, 75 unique alignment positions were accounted for in the top six of the 90 restriction site tags. This corresponds to one percent of reads aligning to restriction enzyme-specific sites in the genomes, with only about five percent of those reads being unique. Unfortunately, this data was not optimal for the genotyping-by-sequencing analysis.

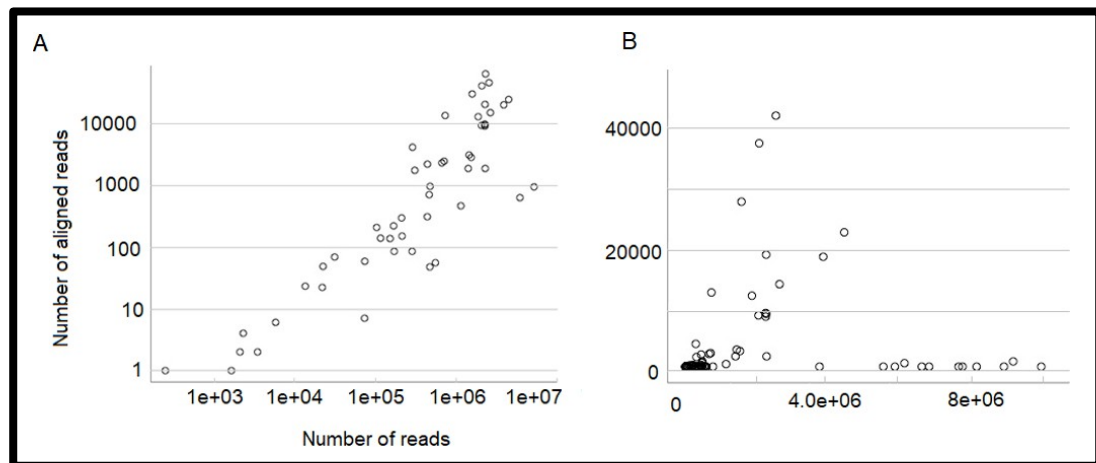


Figure 6.4 Relationship between total read count and aligned read count on logarithmic (A) and simple (B) numerical scales. One dot represents one sample.

By comparing the total number of reads to those which aligned, this one percent becomes clearer. The logarithmic relationship is a linear one, indicating, perhaps intuitively, the more reads from a sample, the more of those reads will map to a reference (Figure 6.4A). Interestingly, when zeros are included in this comparison, we see several samples with high read counts but absolutely no mapping, which perhaps explains the poor recoveries for this analysis. (Figure 6.4B).

6.4 Summary and Conclusions

For the coverage required for this analysis, this is a remarkably and unexpectedly small dataset. There are more diverse reads than would be expected based on the restriction sites used the analysis. These are likely artifacts of degrading DNA, which is to be expected from our historical and dried samples, and unexpected digest of non-target sequences during the analysis itself.

Although usable, this research progression was not continued in favor of more relevant projects in the field, mainly, the development of novel pseudoassemblies and comparison of newly available chlorophyte genomes. In the future, this work may proceed by identifying significant hits and estimating global diversity among those hits.

Chapter 7: FINAL CONCLUSIONS

Along with acknowledgement of the utility of a chlorophyte model organism comes the need to characterize, parse, and report demographic genomic information. Obviously, this field is drawing attention, with three new genomes sequenced and published in last five years (De Clerck et al. 2018, Arimoto et al. 2019, He et al. 2020). This thesis focused on understanding the genomic variation that is characteristic of the genus *Ulva*. We saw it prudent to explore relationships among emerging chlorophyte model organisms, and our primary aim became developing a strategy to do so. The value of this work rests in the comparative analyses and the methodology developed to realize such a broad objective.

7.1 Creation of *de novo* assemblies

Two *de novo* pseudoassemblies, deemed ‘pseudo’ based on substandard genome wide coverage, were created from raw fastq data. The compositional outcome is relatively poor in terms of size, genome coverage, and alignment length. Accordingly, the standard metrics of L50 and N50 are well below generally acceptable ranges in the field. Although the final assemblies hold a relatively small fraction of genomic information, we see their composition as successful in the scope of this project for several reasons.

First, all sequences comprising the assemblies met or exceeded predetermined quality standards. Quality was not sacrificed to artificially inflate continuous read length. While the literature would not qualify the assembly construction as successful based on these parameters, we know that the genetic information in each assembly is reliable and representative. The number of discarded reads may be a downstream effect of tissue processing and the DNA extraction protocol, i.e., the quality of the starting product, not the quality of the data processing.

It is important to remember that at the time, the pseudoassemblies and the reference genome were the only genomic data available. In that context, although the datasets are not as expansive as we may have hoped, we still expect the majority of returned hits to be genomic, non-microbial, and usable. This determination is based on the sampling depth and quality of the assembled reads, in conjunction with the many added steps to the extraction protocol. The analysis that followed was a proof-of-concept exercise to take a more targeted approach to newly emerging chlorophyte genomic data. The results that followed were interesting and warranted their own explanation.

7.2 Characterization of *de novo* pseudoassemblies in the context of established reference genomes

Our analyses of the assemblies report genetic overlap with the reference genome, *U. mutabilis*. Although the number and length of quality-controlled hits to the reference genome may have been less than expected, the agreement between subject and query indicated the target DNA was from the intended organism. Comparison of single nucleotide polymorphisms between Ward and Langston indicate similar levels of variability within the subset of sequences which have direct counterparts in the *U. mutabilis* genome. Additionally, Ward can be conclusively identified as *U. lacinulata* based on the assembly alone, further evidence that the assemblies contain meaningful data (Fort et al. 2020).

The comparison of each pseudoassembly to the reference genome, particularly the dot plot investigative approach used here (Chapter 3), is an exercise in method development used later to compare newly sequenced chlorophytes (Chapter 4). The assemblies do not themselves contain information of significant contribution to the literature. They do, however, provide proof of concept and a repeatable procedure for

proceeding analyses, which has yielded more exciting findings.

In terms of genome structure, Langston and ward look quite different from each other. Ward, however, takes a more standard genomic organization by comparison to the reference genome. The inclusion of singlets is a characteristic of the pseudoassemblies that is not shared by the reference genome. Their inclusion is likely an artifact of the lack of genome wide coverage and unique sequence counts in the raw data files. Ideally a greater number of unique starting sequences would yield higher overlap between groups of sequences, and consequently, arrangement into fewer, larger contigs or scaffolds. There is evidence, however, that connectivity exists between contigs, such that gaps in coverage do not impede us from seeing a higher level of organization among contigs.

Missing from both pseudoassembly dot plots is any indication of a central diagonal identity line, which you would expect to see in closely related organisms. We would absolutely expect to see at least a handful of these matches. Undoubtedly, many of those sequences were not captured by the DNA extraction process or perhaps data processing and assembly construction. There is also the possibility that there are fewer of these matches in actuality than one would expect to find in relatively closely related species within a highly variable genus, given a full coverage genome. The latter hypothesis is the least likely of the three.

A closer look at the match content of Ward against the reference genome shows a few interesting trends. Higher identity hits by percentage represent the longest matches by base length. This tells us there are homologous regions in the two genomes in functionally conserved regions. Three reference genome contigs were identified that matched long (1000+ bases) and high identity (75-100%) hits in the Ward pseudoassembly. Annotations for these matches reveal genes involved in energy and

metabolism and DNA processing, what are considered vital and therefore presumably highly conserved functions. Similarly, three Langston pseudoassembly contigs registered hits on the reference genome. The functions of these genes have more varied functions than the Ward counterparts but are mainly involved in DNA processing and enzyme activity.

The longest Ward contig (19) is a recurring target in each analysis. We can confidently present several findings based on this contig of the Ward pseudoassembly alone. First, the pseudoassembly Ward is of the species *Ulva lacinulata*. A BLAST of contig 19 reveals 97% identity with *U. lacinulata*, a lectotype of *U. lactuca* and a species generally found in more Southern European waters (Hughley et al. 2021). This match characterizes approximately 20% of the entire *U. lacinulata* genome, and as such is an overwhelmingly positive identification.

Second, the *U. mutabilis* published draft genome appears to contain chloroplast DNA, or at the very least, a region of genomic DNA homologous with *U. lacinulata* chloroplast DNA. Ward contig 19, which we can confidently present as *U. lacinulata* chloroplast DNA, matches with greater than 90% identity to *U. mutabilis* contig 178. This reference genome contig matches with 99.9% to the *U. mutabilis* chloroplast genome and covers most of the data files as such (NCBI reference number MK069584/NC_043860.1, 63%). The chloroplast genome of *U. mutabilis* was published separately and the draft genome was intended to only include only nuclear genomic DNA.

Lastly, it is apparent that the density gradient extraction protocol may not be sufficient to fully isolate genomic DNA from chloroplast DNA. DNA extraction for the reference genome was conducted using the CTAB (Cetyltrimethylammonium bromide) protocol, which can isolate total cellular DNA. However, in this particular case, gametes were the extraction target, which notably lack chloroplasts. Therefore, the origin of this contamination and the presence of chloroplast DNA in the reference genome at all is curious (Lutz al. 2011, De Clerck et al. 2018). In a nod to the origin of the organelle, chloroplast DNA gene transfer to nuclear RNA is a documented process, often resulting in homologous regions which are non-functional in the genomic DNA (Ayliffe and Timmis 1992, Huang et al. 2003, Cullis et al. 2008). This is likely the case with the reference genome, since the target organism was gametic and contained no chloroplast DNA. The same cannot be said for the DNA extraction of Ward; either a homologous region with the nuclear DNA has resulted from chloroplast gene transfer or, more simply and perhaps more likely, the extraction protocol is not sufficient to separate genomic DNA from chloroplast DNA.

The comparison of Langston and *U. mutabilis* looks quite different, although it too lacks the characteristic central identity line and identifies repetitive reads in the reference genome. A single high identity match from contig 491 recurs in the majority of reference genome contigs. Among match count and identity, Langston appears to be the more closely related species to the reference genome, with over 50% of matches falling between 80-100% identity, what we would assume to be representative of well conserved genomic regions. This is in comparison to 30% from Ward within the same bin (Figure 3.4A-B). The pattern of highly repetitive sequences is one that is shared between the two pseudoassemblies. It is well established that dot plots are particularly useful in identifying conserved repeating regions in DNA, RNA, and proteins, which

we know to be characteristic of transposable elements (Pontier et al. 2005, Maul et al. 2002, Merski et al. 2020).

We also attempted to identify potential gene functions for those sequences which registered hits to the reference genome but for which there was no documented annotation by exploring close relatives. Most of these examples provide hits to close relative mitochondrial or chloroplast DNA, which explains why there is a lack of gene definition from the reference genome database (Table 3.3). There seems to be a skew towards chloroplast function in the genes that fall into this category. The *U. mutabilis* chloroplast and mitochondrial genomes are published separately, and therefore would not be expected to present hits based on a BLAST search (Melton et al. 2015, Liu al. 2021). We can reasonably conclude that the pseudoassemblies contain not only genomic DNA, but also chloroplast and perhaps mitochondrial DNA, and, again, that the density-dependent, caesium chloride centrifugation procedure was not entirely successful at isolating the target DNA.

7.3 Characterization of newly sequenced chlorophytes

As expected, the comparison between *Ulva* genomes present more similarly to each other as they do to the more distantly related *Caulerpa* assembly. A caveat to these analyses is the lack of whole genome coverage of the two newly sequenced chlorophytes. *U. prolifera* and *C. lentillifera* are missing approximately 25% and 14%, respectively. Some statistics such, as contig number and L50 are not useful to compare between *U. prolifera* and *U. mutabilis*, for example, due to the missing components of the whole genome. It is not entirely intuitive as to why N50 increases and L50 decreases from the reference genome to *U. prolifera*. Lacking coverage and depth from 25% sequence omission would inherently increase the number of contigs as there is less connectivity among sequences. One might expect N50 to decrease, implying

either the larger contigs in the *U. prolifera* assembly dwarf those of the reference genome or it simply takes fewer to reach the 50% genome benchmark because the size of the assembly is smaller. We see clearly that the former is not the case (Figure 4.2).

Although it should not be overlooked that *U. prolifera* may not include upwards of 25% of actual genomic data, there seem to be differences between it and *U. mutabilis* moreso than would be expected based on the statistic alone, as is the example with long terminal repeats (LTR) and long interspersed elements (LIE) (Table 4.1). These two elements are associated with retrotransposon activity, both in a terminal repeating pattern, as the name suggests, and in a non-repeating pattern. Even if the missing 25% of *U. prolifera* were 100% attributed to LTR or LIE, which is extremely unlikely, there would still be quite a gap between the two species. Literature suggests no single automated approach is completely reliable for identifying all transposable elements (Lerat 2010, Lanciano and Crisofari 2020), which may contribute to the differences observed here, as the detection methodologies behind these two assemblies differ (De Clerck et al. 2018, He et al. 2020). More specifically, although both genome assembly processes included utilization of the RepeatMasker v4.07 package, only De Clerck et al. (2018) employed the identification program RepeatModeler 1.0.8 alongside several additional cross-functional pipelines. Hence, the lack of LTRs or LIEs may be a result of the detection efficiency, not a true reflection of the data, and the actual values for these elements are likely closer to those of *U. mutabilis* than those presented by He et al. (2020).

As expected, the comparison of *U. mutabilis* and *U. prolifera* yielded the highest volume of data in this series of analyses. Even so, approximately 90% of the *U. mutabilis* and *U. prolifera* genomes are unaccounted for in the dot plot. We sought to

probe specific conformations of hits in the dot plot, given the scope of data returned from this particular comparison.

The longest continuous and highest identity hits show us that there is a conservation of genome structure, given that those hits fall in discrete locations in either genome. These hits represent the most conserved regions of the alignment, and the biological role of these components appears to be largely centered around protein syntheses and processing, specifically ribosomal RNA genes, which we know to be among the most conserved of all DNA sequences (Eickbush and Eickbush 2007, Isenbarger 2008). Interestingly, but perhaps intuitively, long continuous stretches of higher diversity do not show the same pattern of locational conservation.

A third query of low identity sequences which maintain positional agreement in the genomes showed, in every case, that diversity was attributed to an expansion or deletion event in *U. prolifera*. Inherently, this may not necessarily mean that each low identity hit is itself completely duplicated, triplicated, or deleted when migrating between species and does not consider nested reading frames. This is evidence, however, that there may be specific and rational changes when looking at similar coding regions when comparing the more ancestral *U. mutabilis* to the query. A duplication or triplication from *U. mutabilis* to *U. prolifera* indicates potential gene expansion, while a deletion from *U. mutabilis* to *U. prolifera* is an expansion event in *U. mutabilis*. We also found that transcription factors in these sequences increased or decreased proportionally to the expansion or deletion event. It is possible, however, that the expansion or deletion events correspond to reading frames not dependent upon these transcription sites.

Lastly, we sought examples in which highly conserved genomic regions did not maintain consistency in their location. All examples, selected randomly, were non-

coding RNA genes. ncRNAs are essential and conserved in the toolkit of photosynthetic organisms (Rubin et al. 2015). Within these data we see recurring patterns from previous analyses including conservation of non-coding RNA sequences, an abundance of small hits throughout the comparison which inevitably are filtered from these analyses, and recurring contigs of interest among high identity and more diverse matches (Upro 76.1, Uco 99, and Uco 178). These three particular contigs of interest were found to contain hits for mating loci and, as previously determined, chloroplast DNA (Chapter 3).

It is not uncommon to observe intraspecific variation in chloroplast genomes in land plants, including insertions, deletions, and rearrangement (Park et al. 2020, Park and Oh 2020, Niu et al. 2020). This holds true for our analysis of *Ulva compressa* and *Ulva prolifera*, wherein the most pronounced findings between species were the order of photosystem II protein coding genes, and within species, the order of open reading frames. We know *Ulva* to be robust under irradiation exposure, particularly the proteins of photosystem II (De Clerck et al. 2018, Beer 2022). We refer to *U. mutabilis* here as *U. compressa* to maintain consistency with the referenced publications on which this work is based (Liu and Melton 2021, Chapter 1.4.1, Chapter 4).

Select ORFs appear to be more conserved than others, such that those in *U. compressa* match those in *U. prolifera* with high identity. Others seem to be species-specific. In the most curious scenario, there are *U. compressa* ORFs which map to other ulvophytes, even *Caulerpa*, but not *U. prolifera*. This naturally suggests the evolutionary distance between these two species is consequential, and perhaps selection pressures in the process of speciation are responsible for the counterintuitive nature of these findings.

As expected, the comparison of *U. mutabilis* and *C. lentillifera* offers significantly less data. High identity ($\geq 75\%$), short (< 100 base) hits presumably represent highly conserved sequences, highly repetitive sequences, or both, and comprise approximately half of all sequence matches. Longer and higher diversity ($\leq 25\%$) hits encompass more diverse regions, which fits with the evolutionary distance between these organisms. Aligning the relatively small *C. lentillifera* genome to the larger *U. mutabilis* with large scale, low sensitivity processing yields no hits of rudimentary function for chlorophytes, as would be expected.

7.4 Orthologous groups in sequenced chlorophytes: evolution of green sulfotransferases

These results show green algae do, in fact, possess cell wall and cytosolic sulfatases along with other PAPS-producing enzymes homologs, which is validated by the FTIR curves. There is also evidence in support of elevated SOTs in marine green algal species, both by the proteomic query and FTIR results (Table 5.2, Figure 5.9). In particular, we see expansion in *U. compressa* SOT in both analyses. The current prediction is that these enzymes somehow afford the alga a greater advantage in managing the physical stressors of life in a saline, and often intertidal, environment (Kloareg and Quatrano 1988, Aquino et al. 2005, 2011, Pomin and Mourao 2008, Vishwakarma et al. 2019). This assumption rests in the fact that marine representatives showed the highest proteome hits for CHST and the marine representative's sulphate peak far extended that of the freshwater sample. Of course, this is correlative data and future experiments should more directly explore these results. These results are based on the assumption that more SOTs in *U. compressa* converts to more genes for those enzymes. This is supported by the proteome query but should be confirmed more specifically.

Regarding the orthogroup analyses, there are several interesting findings to be explored. Notably, all species except for our primary target, *U. compressa*, surpassed 80% of genes assigned to orthogroups. We recognize this as a true and reflective interpretation of the data, but we recognize it may also be an artifact. For the former, this would mean that there is truly greater adaptation in *Ulva* relative to other chlorophytes. As such, there are fewer genes in the *U. compressa* genome which have orthologs in other green algal relatives. On the other hand, this reduced value in *U. compressa* could be explained by an over-estimated proteome, such that sampling bias or lack of coverage omit ~7% or more of genes from inclusion in this analysis.

The phylogenies resulting from these analyses present unique relationships between bacterial species and the green algal clades. A single *C. crispus* CHST resolved closest to the oomycete, a fungus-like marine eukaryote, and a mollusc (oyster) (Figure 5.8). Endosymbiosis is likely the explanation between *C. crispus* and the oomycete. A less logical relationship is the oyster and *C. crispus*, although there is evidence of eukaryote-to-eukaryote HGT between the alga *Vaucheria litorea* and sea slugs and as well as the more common prokaryote-to-eukaryote HGT in octopus (Rumpho et al 2008, Conghui et al 2018). Notably, both receptor species in these studies are molluscs. A possible explanation may be independent HGT with similar donor organisms.

Additionally, in rearrangement of the original phylogeny generated from the orthogroup data, the marine bacteria Desulfobulbaceae SOT clusters with *C. crispus*. We know that this particular bacterial sample was collected from a stromatolite metagenome off the coast of South Africa and may represent a P-loop hydrolase rather than a SOT. This enzyme matches the C-terminal end of the SOT, and thus, may provide a false hit with the software. It is peculiar still that it clusters with marine

chlorophytes. A prediction for this arrangement may be either convergent evolution as an adaptive response to a saline environment or exon shuffling. In the latter case, it is possible an HGT event with an Akkermansiaceae-like sequence, followed by replacement of the ‘old’ C-terminal end with the ‘new’ Akkermansiaceae exons. The marine species have an exon structure consistent with this behavior.

The orthogroup analysis provides specific insights into genome evolution in *Ulva*. In following the approach of previous analyses (Chapters 4-5), we choose here to focus on data pertaining to the ulvophytes, *U. compressa* and *C. lentillifera*. There are approximately 20 gene families unique to *Ulva* and *Caulerpa*, of which 6 are candidates for HGT origination. One gene family of this subset (OG0000343) appears to have expanded in *U. compressa*, but not *C. lentillifera*, thus the HGT event must have predated *Ulva* evolution. As it turns out, gene family OG0000343 contains sulfotransferase genes.

For non-*Ulva* sequences, there may not be enough information to build an unambiguous tree. Upon jostling parameters, there are several bacterial SOT sequences that end up closer to the *Chondrus* sequence than to *Ulva*, such as the case of Figure 5.9, in which the marine bacterial Desulfobulbaceae sequence clusters with other marine SOTs in *Chondrus*. Again, these conformations are not stable and will move if tree-building parameters are changed, and thus, they may reflect convergent evolution in response to salinity or other environmental stress, rather than common descent. This complication is not unforeseen, given that we are building trees from bacterial and eukaryotic sequences, which experience differential rates of change.

We present these data and analyses as a contribution to chlorophyte research, with particular impact on the field of comparative genomics. Our findings contribute to the now notion that ulvophytes in particular possess impressive diversity and variation

between and within species. An expanded effort to the genome comparison and SOT work will add even more value to this evolving field.

BIBLIOGRAPHY

- Altenhoff, A.M., Schneider, A., Gonnet, G.H., Dessimoz, C. OMA. 2011. Orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39: D289-294.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
- Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, and Stern, D.L. 2016. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21: 610-617.
- Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev.* 17: 81-92.
- Aquino, R.S., Landeira-Fernandez, A.M., Valente, A.P., Andrade, L.R., and Mourão, P.A.S. 2005. Occurrence of sulfated galactans in marine angiosperms: evolutionary implications. *Glycobiology.* 15: 11-20.
- Aquino, R.S., Grativol, C., and Mourão, P.A.S. 2011. Rising from the Sea: correlations between sulfated polysaccharides and salinity in plants. *PLoS One.* 6: e18862. DOI: 10.1371/journal.pone.0018862.
- Arata, P.X., Alberghina, J., Confalonieri, V., Errea, M.I., Estevez, J.M., and Ciancia, M. 2017. Sulfated Polysaccharides in the Freshwater Green Macroalga *Cladophora surera* Not Linked to Salinity Adaptation. *Front. Plant Sci.* 8: 1927.
- Arieli, A., Sklan, D., & Kissil, G. 1993. A note on the nutritive value of *Ulva lactuca* for ruminants. *Anim. Sci. J.* 57(2): 329-331.

- Ayliffe, M.A. and Timmis, J.N. 1992. Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. *Theoret. Appl. Genetics*. 85: 229-238.
- Beer, S. Photosynthetic traits of the ubiquitous and prolific macroalga *Ulva* (Chlorophyta): a review. 2022. *Eur. J. Phycol.* DOI: 10.1080/09670262.2022.2150894.
- Beer, S., Israel, A., Drechsler, Z., and Cohen, Y. 1990. Photosynthesis in *Ulva fasciata*: Evidence for an Inorganic Carbon Concentrating System, and Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase CO(2) Kinetics. *Plant Physiol.* 94(4): 1542-1546.
- Bittick, S.J., Clausing, R.J., Fong, C.R., Scoma, S.R., and Fong, P. 2019. A rapidly expanding macroalga acts as a foundational species providing trophic support and habitat in the South Pacific. *Ecosystems*. 22: 165-173.
- Bliding, C. 1968. A critical survey of European taxa *Ulvales* II. *Ulva*, *Ulvaria*, *Monostroma*, *Kornmannia*. *Bot. Not.* 121:535-629.
- Blomme, J., Liu, X., Jacobs, T.B., and De Clerck, O. 2021. A molecular toolkit for the green seaweed *Ulva mutabilis*. *Plant physiol.* 186(3): 1442-1454.
- Blomster, J., Bäck, S., Fewer, D.P., Kiirikki, M., Lehvo, A., Maggs, C.A., and Stanhope, M.J. 2002. Novel morphology in *Enteromorpha* (Ulvophyceae) forming green tides. *Am. J. Bot.* 89(11): 1756-1763.
- Blomster, J., Maggs, C.A., and Stanhope, M.J. 1998. Molecular and morphological analysis of *Enteromorpha intestinalis* and *E. compressa* (Chlorophyta) in the British Isles. *J. Phycol.* 34: 319-340.

- Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*. 30(15). 2114-2120.
- Bonen, L. and Vogel, J. 2001. The ins and outs of group II introns. *Trends Genet.* 17(6): 322-331.
- Brodie, J. and Lewis, J. 2007. Unravelling the algae. London, U.K.
- Bruhn, A., Dahl, J., Nielsen, H.B., Nikolaisen, L., Rasmussen, M.B., Markager, S., Olesen, B., Arias, C., and Jensen, P.D. 2011. Bioenergy potential of *Ulva lactuca*: Biomass yield, methane production and combustion. *Bioresour. Technol.* 102(3): 2595-2604.
- Bryhni, E. 1974. Genetic control of morphogenesis in the multicellular alga *Ulva mutabilis*. *Dev. Biol.* 37: 273-279.
- Bubrick, P. 1991. Production of astaxanthin from *Haematococcus*. *Bioresour. Technol.* 38: 237-239.
- Buchfink, B., Xie, C., and Huson D.H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* 12(1): 59-60.
- Burrows, E.M. 1991. Seaweeds of the British Isles: Volume 2, Chlorophyta. Natural History Museum, London. p. 222.
- Bushnell, B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-7065E. Retrieved from <https://escholarship.org/uc/item/1h3515gn>.
- Cabanettes, F. and Klopp, C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. DOI: 10.7717/peerj.4958.

- Camargo, J.A. and Alonso, A. 2006. Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: A global assessment. *Environ. Int.* 32(6): 831-849.
- Canfield, D.E. and Farquhar, J. 2009. Animal evolution, bioturbation, and the sulfate concentration of the oceans. *P. Natl. A. Sci.* 106(20): 8123-7.
- Carr, S.M. and Griffiths, O.M. 1987. Preparative density-gradient ultracentrifugation of DNA. *Biochem. Genet.* 25: 385-390.
- Cao, Y., Wu, Y., Fang, Z., Cui, X., Liang, J., and Song, X. 2019. Spatiotemporal Patterns and Morphological Characteristics of *Ulva prolifera* Distribution in the Yellow Sea, China in 2016-2018. *Remote Sens.* 11(4): 445-472.
- Chae, H., Kim, E.J., Lim, S., Kim, H.S., Choi, H., Kim, S., and Kim, J.H. 2021. Morphology and phylogenetic relationships of two new Antarctic species in phylogroup *Chloromonadinia* (Volvocales, Chlorophyceae), *Phycologia*, 60(3): 225-236.
- Chapman, V. 1950. Seaweeds and their uses. London, U.K.
- Chbani, A., Majed, S., Mawlawi, H., and Kammoun, M. 2015. The use of seaweed as a bio-fertilizer; Does it influence Proline and Chlorophyll concentration in plants treated? *AJMAP*. 1(1): 67-77.
- Chekroun, K.B., Sánchez, E., and Baghour, M. 2014. The role of algae in bioremediation of organic pollutants. *Int. Res. J. Public Environ. Health*. 1(2): 19-32.

- Chen, F., Mackey A.J., Stoeckert, C.J., Roos, D.S. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34: D363-368.
- Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C.Y., Ecker, J.R., Cantu, D., Rank, D.R., and Schatz, M.C. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods.* 13(12): 1050-1054.
- Ciancia, M., Alberghina, J., Arata, P.X., Benavides, H., Leliaert, F., Verbruggen, H., and Estevez, J.M. 2012. Characterization of cell wall polysaccharides of the coencocytic green seaweed *Bryopsis plumose* (Bryopsidaceae, Chlorophyta) from the Argentine coast. *J. Phycol.* 48(2): 326-335.
- Ciancia, M., Fernández, P.V., and Leliaert, F. 2020. Diversity of Sulfated Polysaccharides From Cell Walls of Coenocytic Green Algae and Their Structural Relationships in View of Green Algal Evolution. *Front. Plant Sci.* 11:554585. DOI: 10.3389/fpls.2020.554585.

Cock, M.J., Lieven, S., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.M., Badger, J.H., Beszteri, B., Billiau, K., Bonnet, E., Bothwell, J.H., Bowler, C., Boyen, C., Brownlee, C., Carrano, C.J., Charrier, B., 1,2, Cho, G.Y., Coelho, S.M., Collén, J., Corre, E., Da Silva, C., Delage, L., Delaroque, N., Dittami, S.M., 1,2, Doulbeau, S., Elias, M., Farnham, G., Gachon, C.M.M., Gschloess, B., Heesch, S., Jabbari, K., Jubin, C., Kawai, H., Kimura, K., Kloareg, B., Küpper, F.C., Lang, D., Bail, A.L., Leblanc, C., Lerouge, P., Lohr, M., Lopez, P.J., Martens, C., Maumus, F., Michel, G., Miranda-Saavedra, D., Morales, J., Moreau, H., Motomura, T., Nagasato, C., Napoli, C.A., Nelson, D.R., Nyvall-Colle, P., Peters, A.F., Pommier, C., Potin, P., Poulain, J., Quesneville, H., Read, B., Rensing, S.A., Ritter, A., Rousvoal, S., Samanta, M., Samson, G., Schroeder, D.C., Ségurens, B., Strittmatter, M., Tonon, T., Tregear, J.W., Valentin, K., Dassow, P.V., Yamagishi, T., Van de Peer, Y., Wincker, P. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nat.* 465: 617-621. DOI:10.1038/nature09016.

Cock, P.J.A., Bonfield, J., Chevreux, B. and Li, H. 2015. SAM/BAM format v1.5 extensions for *de novo* assemblies. *BioRxiv*. DOI:10.1101/020024.

Cocquyt, E., Verbruggen, H., Leliart, F., and De Clerck, O. 2010. Evolution and cytological diversification of the green seaweeds (Ulvophyceae). *Mol. Biol. Evol.* 27(9): 2052-2061.

Conghui, L., Bo, L., Yan, Z., Fan, J., Yuwei, R., Shuqu, L., Hengchao, W., and Wei, F. 2018. Data on horizontally transferred genes in California two-spot octopus, *Octopus bimaculoides*. *Data in brief*. 19:1274-1286. DOI: 10.1016/j.dib.2018.05.132.

- Cordeiro, L. M., Carbonero, E. R., Sassaki, G. L., Reis, R. A., Stocker-Worgotter, E., Gorin, P. A. J., and Iacomini, M. 2006. A fungus-like β -galactofuranan in the cultivated *Trebouxia* photobiont of the lichen *Ramalina gracilis*. *FEMS Microbiol. Lett.* 244: 193-198.
- Cosenza, V.A., Navarro, D.A., Ponce, N.M.A., Stortz, C.A. 2017. Seaweed Polysaccharides: Structure and Applications. Industrial Applications of Renewable Biomass Products, ed. Springer. 75-116.
- Cruz-Suárez, L.E., León, A., Peña-Rodríguez, A., Rodríguez-Peña, G., Moll, B., and Ricque-Marie, D. 2010. Shrimp/*Ulva* co-culture: A sustainable alternative to diminish the need for artificial feed and improve shrimp quality. *Aquacult.* 301: 64-68.
- Cullis, C. A., Vorster, B. J., Van Der Vyver, C., and Kunert, K. J. 2009. Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants? *Ann. Bot.* 103(4): 625-633.
- Dang, T.T., Yasufumi, M., and Dang, D.K. 2012. Potential of *Ulva sp.* In biofiltration and bioenergy production. *J. Viet. Env.* 3(1): 55-59.
- Davey, J.W. and Blaxter, M.L. 2011. RADseq: next-generation population genetics. *Brief. Funct. Genom.* 9(5): 416-423.

- De Clerck, O., Kao, S., Bogaert, K.A., Blomme, J., Foflonker, F., Kwantes, M., Vancaester, E., Vanderstraeten, L., Aydogdu, E., Boesger, J., Califano, G., Charrier, B., Clewes, R., Del Cortona, A., D'Hondt, S., Fernandez-Pozo, N., Gachon, C.M., Hanikenne, M., Lattermann, L., Leliaert, F., Liu, X., Maggs, C.A., Popper, Z.A., Raven, J.A., Van Bel, M., Wilhelmsson, P.K.I., Bhattacharya, D., Coates, J.C., Rensing, S.A., Straeten, D.V.D., Vardi, A., Sterck, L., Vandepoele, K., de Peer, Y.V., Wichard, T., and Bothwell, J.H. 2018. Insights into the evolution of multicellularity from the sea lettuce genome. *Curr. Biol.* 28: 2921-2933.
- Del Cortona, A., Jackson, C.J., Bucchini, F., Van Bel, M., D'hondt, S., Skaloud, P., Delwiche, C.F., Knoll, A.H., Raven, J.A., Verbruggen, H., Vandepoele, K., De Clerck, O., and Leliaert, F. 2020. Neoproterozoic origin and multiple transitions to macroscopic growth in green seaweeds. *P. Natl. A. Sci. USA.* 117 (5): 2551-2559.
- Demirbas, A. 2010. Use of algae as biofuel sources. *Energy Convers. Manag.* 51(12): 2738-2749.
- Deschamps, S., Llaca, V., and May, G.D. 2012. Genotyping-by-sequencing in plants. *Biol.* 1: 460-483.
- Dewinta, A.F., Halomoan, Y.T., Susetya, I.E., Yusni, E., and Fafhilah, A. 2021. Diversity and cover percents of macroalgae species in intertidal zone of Pane. *Earth. Envi. Sci.* 782: 042008.
- Diaz, R.J. and Rosenberg, R. 1995. Marine benthic hypoxia: a review of its ecological effects and the behavioural responses of benthic macrofauna. *Oceanogr. Mar. Biol.* 33: 245-303.

- Dion, P., De Reviers, B., and Coat, G. 1998. *Ulva armoricana* sp. nov. (Ulvales, Chlorophyta) from the coasts of Brittany (France). I. Morphological identification. *Eur. J. Phycol.* 33(1): 73-80.
- Domozych, D.S., Ciancia, M., Fangel, J.U., Mikkelsen, M.D., Ulvskov, P. and Willats, W.G.T. 2012. The cell walls of green algae: a journey through evolution and diversity. *Front. Plant Sci.* 3(82).
- Draisma, S.G., van Reine, W.F.P., Sauvage, T., Belton, G.S., Gurgel, C.F.D., Lim, P. E. and Phang, S.M. 2014. A re-assessment of the infra-generic classification of the genus *Caulerpa* (Caulerpaceae, Chlorophyta) inferred from a time-calibrated molecular phylogeny. *J. Phycol.* 50: 1020-1034.
- Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Zhao, X., Wang, J., Liu, K., Qin, P., Yang, X., Zhu, L., Li, S., and Liang, C. 2017. Sequencing and de novo assembly of a near complete indica rice genome. *Nat Commun.* 8: 15324.
- Dunker, S. and Wilhelm, C. 2018. Cell wall structure of coccoid green algae as an important trade-off between biotic interference mechanisms and multidimensional cell growth. *Front. Microbiol.* DOI: 10.3389/fmicb.2018.00719.
- Eickbush, T.H. and Eickbush D.G. 2007. Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. *Genetics.* 175(2): 477-485.
- Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotype-by-sequencing (GBS) approach for high diversity species. *Plos One.* 6(5): e19379. DOI: 10.1371/journal.pone.0019379.

- Emms, D.M., Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16: 157.
- Emms, D.M., Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20: 238.
- Estevez, J. M., Leonardi, P. I., and Alberghina, J. S. 2008. Cell wall carbohydrate epitopes in the green alga *Oedogonium bharuchae* F. minor (Oedogoniales, Chlorophyta). *J. Phycol.* 44: 1257-1268.
- Estevez, J.M., Fernandez, P.V., Kasulin, L., Dupree, P., Ciancia, M. 2009. Chemical and in situ characterization of macromolecular components of the cell walls from the green seaweed *Codium fragile*. *Glycobiology.* 19(3): 212-228.
- Ergün, S., Soyutürk, M., Güroy, B., Güroy, D., and Merrifield, D. 2009. Influence of Ulva meal on growth, feed utilization, and body composition of juvenile Nile tilapia (*Oreochromis niloticus*) at two levels of dietary lipid. *Aquacult. Int.* 17: 355-361.
- Fang, L., Leliaert, F., Zhang, Z.-H., Penny, D. and Zhong, B.-J. 2017. Evolution of the Chlorophyta: Insights from chloroplast phylogenomic analyses. *J. Syst. Evol.* 55: 322-332.
- Fernandez-Garcia, C., Riosmena-Rodriguez, R., Wysor, B., Tejada, O.L., and Cortes, J. 2011. Checklist of the Pacific marine macroalgae of Central America. *Bot. Mar.* 54(1): 53-73.
- Fjeld, A. 1972. Genetic control of cellular differentiation in *Ulva mutabilis*. *Dev. Biol.* 28: 326-343.

- Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H. Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A. L. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids*. 45(D1): D190-D199.
- Fischer, M.C., Rellstab, C., Tedder, A., Zoller, S., Gugerli, F., Shimizu, K.K., Holderegger, R., and Widmer, A. 2013. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molec. Ecol.* 22: 5594-5607.
- Fleurence, J. 1999. Seaweed proteins: biochemical, nutritional aspects and potential uses. *Trends Food Sci. Technol.* 10: 25-28.
- Foissac, S., Gouzy, J., Rombauts, S., Mathe, C., Amselem, J, Sterck, L., de Peer, Y.V., Rouze, P., and Schiex, T. 2008. Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* 3(2): 87-97.
- Fletcher, R.L. 1996. The Occurrence of “Green Tides”- a Review. In: Schramm W., Nienhuis P.H. (eds) Marine Benthic Vegetation. Ecological Studies (Analysis and Synthesis), vol 123. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-61398-2_2.

- Fort, A., McHale, M., Cascella, K., Potin, P., Usadel, B., Guiry, M.D., and Sulpice, R. 2021. Foliose *Ulva* Species Show Considerable Inter-Specific Genetic Diversity, Low Intra-Specific Genetic Variation, and the Rare Occurrence of Inter-Specific Hybrids in the Wild. *J. Phycol.* 57(1): 219-233.
- Garnier, J., Riou, P., Le Gendre, R., Ramarson, A., Billen, G., Cugier, P., Schapira, M., Théry, S., Thieu, V., and Ménesguen, A. 2019. Managing the Agri-Food System of Watersheds to Combat Coastal Eutrophication: A Land-to-Sea Modelling Approach to the French Coastal English Channel. *Geosciences*. 9(10): 441-467.
- Garrett, C.W. 1992. On global climate change, carbon dioxide, and fossil fuel combustion. *Prog. Energ. Combust.* 18(5): 369-407.
- Ghaderiardakani, F., Coates, J., and Wichard, T. 2017. The cross-kingdom interaction in the marine macroalgae *Ulva*: revisiting the lottery theory. *Phycologia*. 56(4): 62.
- Ghaderiardakani, F., Coates, J.C., and Wichard, T. 2017. Bacteria-induced morphogenesis of *Ulva intestinalis* and *Ulva mutabilis* (Chlorophyta): a contribution to the lottery theory. *FEMS Microbiol. Ecol.* 93(8): fix094.
- Gibbs, S.P. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.* 56(22): 2883-2889.
- Gish, W. and States, D.J. 1993. Identification of protein coding regions by database similarity search. *Nature Genet.* 3:266-272.
- Guidone, M., Thornber, C., Wysor, B., and O’Kelly, C.J. 2103. Molecular and morphological diversity of Narragansett Bay (RI, USA) *Ulva* (Ulvales, Chlorophyta) populations. *J. Phycol.* 49: 979-995.

- Gladyshev, M.I. and Gubelit, Y.I. 2019. Green Tides: New Consequences of the Eutrophication of Natural Waters (Invited Review). *Contemp. Probl. Ecol.* 12: 109-125.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids.* 40(D1): D1178-D1186.
- Hayden, H.S. and Waaland, J.R. 2002. Phylogenetic systematics of the Ulvaceae (Ulvales, Ulvophyceae) using chloroplast and nuclear DNA sequences. *J. Phycol.* 38: 1200-1212.
- Hayden, H.S., Blomster, J., Maggs, C.A., Silva, P.C., Stanhope, M.J., and Waaland, J.R. 2003. Linnaeus was right all along: *Ulva* and *Enteromorpha* are not distinct genera. *Eur. J. Phycol.* 38: 277-294.
- Hayden, H.S. and Waaland, J.R. 2004. A molecular systematic study of *Ulva* (Ulvaceae, Ulvales) from the northeast Pacific. *Phycologia.* 43(4): 364-382.
- He, Y., Shen, S., Yu, D., Wang, Y., Yin, J., Wang, Z., and Ye, Y. 2021. The *Ulva prolifera* genome reveals the mechanism of green tides. *J. Oceanol Limnol.* 39: 1458-1470.
- Hebert, P.N.D., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. *R. Soc.* 270: 313-321.
- Hiraoka, M. and Enomoto, S. 1998. The induction of reproductive cell formation of *Ulva perusa* Kjellman (Ulvales, Ulvophyceae). *Phycol. Res.* 46: 199-203.

- Hiraoka, M., Dan, A., Shimada, S., Hagihira, M., Migita, M., and Ohno, M. 2003. Different life histories of *Enteromorpha prolifera* (Ulvales, Chlorophyta) from four rivers on Shikoku Island, Japan. *Phycologia*. 42(3): 275-284.
- Hiraoka, M., Ohno, M., Kawaguchi, S., and Yoshida, G. 2004. Crossing test among gloating *Ulva* thalli forming 'green tide' in Japan. *Hydrobiologia*. 512: 239-45.
- Hiraoka, M., Ichihara, K., Zhu, W., Ma, J., and Shimada, S. 2011. Culture and hybridization experiments on an *Ulva* clade including the Qingdao strain blooming in the Yellow Sea. *Plos One*. 6(5): e19371. DOI: 10.1371/journal.pone.0019371.
- Hirschmann, F., Krause, F., and Papenbrock, J. 2014. The multi-protein family of sulfotransferases in plants: composition, occurrence, substrate specificity, and functions. *Front. Plant Sci*. 5:556.
- Hoeksema, B. W. and van den Hoek, C. 1983. The Taxonomy of *Ulva* (Chlorophyceae) from the Coastal Region of Roscoff (Brittany, France). *Bot. Mar.* 26(2): 65-86.
- Hofmann, L.C., Nettleton, J.C., Neefus, C.D., and Mathieson, A.C. 2010. Cryptic diversity of *Ulva* (Ulvales, Chlorophyta) in the Great Bay Estuarine System (Atlantic USA): introduced and indigenous distromatic species. *Eur. J. Phycol.* 45(3): 230-239.
- Holzinger, A., Herburger, K., Kaplan, F., and Lewis, L.A. 2015. Dessication tolerance in the chlorophyte green algae *Ulva compressa*: does cell wall architecture contribute to ecological success? *Planta*. 242: 477-492.

- Holzinger, A., Herburger, K., Blaas, K., Lewis, L.A., and Karsten, U. 2017. The terrestrial green macroalga *Prasiola calophylla* (Trebouxiophyceae, Chlorophyta): ecophysiological performance under water-limiting conditions. *Protoplasma*. 254: 1755-1767.
- Hong, D.D., Hien, H.M., and Son, P.N. 2007. Seaweeds from Vietnam used for functional food, medicine, and biofertilizer. *J. Appl. Phycol.* 19: 817-826.
- Hoxmark, R.C. 1975. Experimental analysis of the life cycle of *Ulva mutabilis*. *Bot. Mar.* 18: 123-129.
- Hu, L., Hu, C., and Ming-Xia, H.E. 2017. Remote estimation of biomass of *Ulva prolifera* macroalgae in the Yellow Sea. *Remote Sens. Environ.* 192: 217-227.
- Huang, C.Y., Ayliffe, M.A., and Timmis, J.N. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature*. 422(6927):72-76.
- Hughley, J.R., Gabrielson, P.W., Maggs, C.A., and Mineur, F. 2021. Genomic analysis of the lectotype specimens of European *Ulva rigida* and *Ulva lacinulata* (Ulvaceae, Chlorophyta) reveals the ongoing misapplication of names. *J. Phycol.* DOI: 10.1080/09670262.2021.1914862.
- Huston, M.A. and Wolverton, S. 2009. The global distribution of net primary production: resolving the paradox. *Ecol. Monogr.* 79(3): 343-377.
- Isenbarger, T.A., Carr, C.E., Johnson, S.S., Finney, M., Church, G.M., Gilbert, W., Zuber, M.T., and Ruvkun, G. 2008. The Most Conserved Genome Segments for Life Detection on Earth and Other Planets. *Astrobiology*. DOI 10.1007/s11084-008-9148-z.

- Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antiniou, E., McCombie, W.R., Presting, G.G., McMullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., and Ware, D. 2017. Improved maize reference genome with single-molecule technologies. *Nature*. 546:524-552.
- Kakuta, Y., Pedersen, L.G., Pedersen, L.C., Negishi, M. 1998. Conserved structural motifs in the sulfotransferase family. *Trends Biochem. Sci.* 23(4): 129-130.
- Kakuta, Y., Sueyoshi, T., Negishi, M., Pedersen, L.C. 1999. Crystal structure of the sulfotransferase domain of human heparan sulfate N-deacetylase/ N-sulfotransferase 1. *J. Biol. Chem.* 274(16): 10673-10676.
- Katsaros, C., Weiss, A., Llangos, I., Theodorou, I. and Wichard, T. 2017. Cell structure and microtubule organization during gametogenesis of *Ulva mutabilis* Foyn (Chlorophyta). *Bot. Mar.* 60(2): 123-135.
- Kenny, P. and Flynn, K.J. 2017. Physiology limits commercially viable photoautotrophic production of microalgal biofuels. *J. Appl. Phycol.* 29: 2713-2727.
- Kerswell, A.P. 2006. Global biodiversity patterns of benthic marine algae. *Ecology*. 87(10): 2479-2488.
- Kirk, D.L. 1998. *Volvox*. New York: Cambridge University Press.
- Khan, A.M., Fatima, M., Hussain, M.S., and Yasmeen, K. 2016. Biodiesel production from green seaweed *Ulva fasciata* catalyzed by novel waste catalysts from Pakistan Steel Industry. *Chin. J. Chem. Eng.* 24(8): 1080-1086.

- Kloareg, B. and Quatrano, R.S. 1988. Structure of the Cell Walls of Marine Algae and Ecophysiological Functions of the Matrix Polysaccharides. *Oceanogr. Mar. Biol.* 26: 259-315.
- Koemann, R.P.T. and van den Hoek, C. 1981. The taxonomy of *Ulva* (Chlorophyceae) in the Netherlands. *Brit. Phycol. J.* 16(1): 9-53.
- Kotrbaček, V., Doubek, J., and Doucha, J. 2015. The chlorococcalean alga *Chlorella* in animal nutrition: a review. *J. Appl. Phycol.* 27: 2173-2180.
- Krupnik, N., Rinkevich, B., Paz, G., Douek, J., Lewinsohn, E., Israel, A., Carmel, N., Mineur, F., and Maggs, C.A. 2018. Native, invasive and cryptogenic *Ulva* species from the Israeli Mediterranean Sea: risk and potential. *Medit. Mar. Sci.* 19(1): 132-146.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35(6): 1547–1549.
- Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., and Ussery, D.W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35(9): 3100-3108.
- Lahaye, M. and Robic, A. 2007. Structure and functional properties of ulvan, a polysaccharide from green seaweeds. *Biomacromolecules.* 8(6): 1765-1774.
- Lanciano, S. and Cristofari, G. 2020. Measuring and interpreting transposable element expression. *Nat. Rev. Genet.* 21: 721-736.

- Lawton, R., Mata, L., de Nys, R., and Paul, N.A. 2013. Algal bioremediation of waste waters from land-based aquaculture using *Ulva*: selecting target species and strains. *Plos One*. 8(10): e77344. DOI:10.1371/journal.pone.0077344.
- Lefebvre, A. and Devreker, D. 2020. First Comprehensive Quantitative Multi-Parameter Assessment of the Eutrophication Status from Coastal to Marine French Waters in the English Channel, the Celtic Sea, the Bay of Biscay, and the Mediterranean Sea. *J. Mar. Sci. Eng.* 8: 561-585.
- Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., and De Clerck, O. 2012. Phylogeny and molecular evolution of the green algae. *Crit. Rev. Plant. Sci.* 31: 1-46.
- Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*. 104: 520-533.
- Lewis, L.A. 2007. Chlorophyta on land. Seckbach J. (eds) *Algae and Cyanobacteria in Extreme Environments. Cellular Origin, Life in Extreme Habitats and Astrobiology*, vol 11. Springer, Dordrecht. DOI: 10.1007/978-1-4020-6112-7_31.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*. 25(16): 2078-2079.
- Lin, H. and Qin, S. 2014. Tipping points in seaweed genetic engineering: scaling up opportunities in the next decade. *Mar. Drugs*. 12(5): 3025-3045.

- Liu, D., Keesing, J.K., He, P., Wang, Z., Shi, Y., and Wang, Y. 2013. The world's largest macroalgal bloom in the Yellow Sea, China: Formation and implications. *Estuar. Coast. Shelf Sci.* 129: 2-10.
- Loughnane, C.J., McIvor, L.M., Rindi, F., Stengel, D.B., and Guiry, M.D. 2008. Morphology, *rbcL* phylogeny and distribution of distromatic *Ulva* (Ulvophyceae, Chlorophyta) in Ireland and southern Britain. *Phycologia*. 47(4): 416-429.
- Lovlie, A. 1978. On the genetic control of cell cycles during morphogenesis in *Ulva mutabilis*. *Dev. Biol.* 64: 164-177.
- Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Fioretto, L.D.R., Davidson, C., Dodiya, K., Houdaigui, B.E., Fatima, R., Gall, A., Giron, C.G., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O.G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J.G., Marugán, J.C., Maurel, T., McMahon, A.C., Mohanan, S., Moore, B., Muffato, M., Oheh, D.N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M.P., Salam, A.I.A., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., Silva, N.D., Flint, B., Frankish, A., Hunt, S.E., Ilesley, G.R., Langridge, N., Loveland, J.E., Martin, F.J., Mudge, J.M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S.J., Cunningham, F., Yates, A.D., Zerbino, D.R., Flicek, P. 2021. Ensembl. *Nucleic Acids*. 49(1): 884-891.

- Luo, M.B., Lui, F., and Xu, Z.L. 2012. Growth and nutrient uptake capacity of two co-occurring species, *Ulva prolifera* and *Ulva linza*. *Aquat. Bot.* 100: 18-24.
- Lutz, K.A., Wang, W., Zdepski, A., and Michael, T.P. 2011. Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnol.* 11: 54. DOI: 10.1186/1472-6750-11-54.
- Lyngby, J.E., Mortensen, S., Ahrensberg, N. 1999. Bioassessment Techniques for Monitoring of Eutrophication and Nutrient Limitation in Coastal Ecosystems. *Mar. Pollut. Bull.* 39: 212-223.
- Machado, L., Kinley, R.D., Magnusson, M., de Nys, R., and Tomkins, N.W. 2015. The potential of macroalgae for beef production systems in Northern Australia. *J Appl Phycol.* 27: 2001-2005.
- Mahendran, S. and Saravanan, S. 2017. Molecular taxonomy of green seaweeds *Ulva lactuca* and *Caulerpa taxifolia* through phylogenetic analysis. *Indian J. Mar. Sci.* 46(2): 414-9.
- Mainreck, N., Brézillon, S., Sockalingum, G.D., Maquart, F.X., Manfait, M., and Wegrowski, Y. 2011. Rapid characterization of glycosaminoglycans using a combined approach by infrared and Raman microspectroscopies. *J. Pharm. Sci.* 100(2): 441-50.
- Malta, E.J., Draisma, S.G.A., and Kamermans, P. 1999. Freefloating *Ulva* in the southwest Netherlands: species or morphotypes? A morphological, molecular and ecological comparison. *European Journal of Phycology.* 34: 443-54.
- Markham, J.W. and Hagmeier, E. 1982. Observations on the effects of germanium dioxide on the growth of macro-algae and diatoms. *Phycologia* 21(2): 125-30.

- Matsui, T. and Kioke, Y. 2010. Methane fermentation of a mixture of seaweed and milk at a pilot-scale plant. *J. Biosci. Bioeng.* 110(5): 558-63.
- Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H., and Stern, D.B. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell.* 14(11): 2659-79.
- Melton, J.T. III, Leliaert, F., Tronholm, A., and Lopez-Bautista, J.M. 2015. The Complete Chloroplast and Mitochondrial Genomes of the Green Macroalga *Ulva* sp. UNA00071828 (Ulvophyceae, Chlorophyta). *PLoS One.* 10(4): e0121020.
- Merski, M., Mlynarczyk, K., Ludwiczak, J., Skrzeczkowski, J., Dunin-Horkawicz, S., and Gorna, M.W. 2020. Self-analysis of repeat proteins reveals evolutionarily conserved patterns. *BMC Bioinform.* 21: 179-96.
- Milledge, J.J., Smith, B., Dyer, P.W., and Harvey, P. 2014. Macroalgae-Derived Biofuel: A Review of Methods of Energy Extraction from Seaweed Biomass. *Energies.* 7: 7194-222.
- Mohamed, S., Hashim, S.N., and Rahman, H.A. 2012. Seaweeds: A sustainable functional food for complementary and alternative therapy. *Trends Food Sci. Technol.* 23: 83-96.
- Montzka, S., Dlugokencky, E., and Butler, J. 2011. Non-CO2 greenhouse gases and climate change. *Nature.* 476: 43-50.
- Moreira, A., Cruz, S., Marques, R., and Cartaxana, P. 2021. The underexplored potential of green macroalgae in aquaculture. *Rev Aquacult.* 14: 5-26.

- Nagarajan, N. and Pop, M. 2013. Sequence assembly demystified. *Nat. Rev. Genet.* 14: 147-67.
- Nakamura, Y., Sasaki, N., Kobayashi, M., Ojima, N., Yasuike, M., Shigenobu, Y., Satomi, M., Fukuma, Y., Shiwaku, K., Tsujimoto, A., Kobayashi, T., Nakayama, I., Ito, F., Nakajima, K., Sano, M., Wada, T., Kuhara, S., Inouye, K., Gojobori, T., and Ikeo, K. 2013. The first symbiont-free genome sequence of marine alga, Susabi-nori (*Pyropia yezoensis*). *PLoS One.* 8(3): e57122.
- Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R., and Hohenlohe, P.A. 2013. Genotyping-by-sequencing in ecological and conservation genomics. *Molec. Ecol.* 22: 2841-7.
- Niu, E., Jiang, C., Wang, W., Zhang, Y., Zhu, S. 2020. Chloroplast Genome Variation and Evolutionary Analysis of *Olea europaea* L. *Genes.* 11(8): 879.
- Novikova, O. and Belfort, M. 2017. Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends in Genet.* 33(11): 773-83.
- Oertel, W., Wichard, T., and Weissgerber, A. 2015. Transformation of *Ulva mutabilis* (Chlorophyta) by vector plasmids integrating into the genome. *J. Phycol.* 51: 963-79.
- O'Rourke, C., Gregson, T., Murray, L., Sadler, I.H., and Fry, S.C. 2015. Sugar composition of the pectic polysaccharides of charophytes, the closest algal relatives of land-plants: presence of 3-O-methyl-D-galactose residues. *An. Bot.* 116: 225-36.

- Ortega, A., Geraldi, N.R., Alam, I., Kamau, A.A., Acinas, S.G., Logares, R., Gasol, J.M., Massana, R., Krause-Jensen, D., and Duarte, C.M. 2019. Important contribution of macroalgae to oceanic carbon sequestration. *Nat. Geosci.* 12: 748-54.
- Paliwal, R., Gusain, P, Kaphaliya, B., and Uniyal, S. 2018. Algal potential for inorganic and organic pollutants decontamination. *Water Biology*, ed. Khanna and Bhutiani (New Delhi: Discovery Publishing House). 309-317.
- Park, J. and Oh, S.H. 2020. A second complete chloroplast genome sequence of *Fagus Multinervis* Nakai (Fagaceae): intraspecific variations on chloroplast genome. *Mitochondr. DNA.* 5(2): 1868-9.
- Park, J., Xi, H., and Kim, Y. 2020. The Complete Chloroplast Genome of *Arabidopsis thaliana* Isolated in Korea (Brassicaceae): An Investigation of Intraspecific Variations of the Chloroplast Genome of Korean *A. thaliana*. *Int. J. Genomics*. DOI: 10.1155/2020/3236461.
- Percival, E. 1979. The polysaccharides of green, red and brown seaweeds: their basic structure, biosynthesis and function. *Brit. J. Phycol.* 14(2):103-17.
- Petrotschenko, E. V., Pedersen, L. C., Borchers, C. H., Tomer, K.B., and Negishi, M. 2001. The dimerization motif of cytosolic sulfotransferases. *FEBS Lett.* 490: 39-43.
- Phillips, J. A. 1990. Life History Studies of *Ulva rigida* C. Ag. and *Ulva stenophylla* S. et G. (Ulvaceae, Chlorophyta) in Southern Australia. *Bot. Mar.* 33(1): 79-84.
- Pirian, K., Piri, K., Sohrabipour, J., Jahromi, S.T., and Blomster, J. 2016. Molecular and morphological characterization of *Ulva chaugulii*, *U. paschima* and *U. ohnoi* (Ulvophyceae) from the Persian Gulf, Iran. *Bot. Mar.* 59: 147-58.

- Pomin, V. H. and Mourao, P. A. 2008. Structure, biology, evolution, and medical importance of sulfated fucans and galactans. *Glycobiology*. 18: 1016-27.
- Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M., Lerbs-Mache, S., Colot, V., and Lagrange, T. 2005. Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in *Arabidopsis*. 19: 2030-40.
- Racine, P., Marley, A., Froehlich, H.E., Gaines, S.D., Ladner, I., MacAdam-Somer, I., and Bradley, D. 2021. A case for seaweed aquaculture inclusion in U.S. nutrient pollution management. *Mar. Policy*. 129: 104506.
- Ratana-Arporn, P. and Chirapart, A. 2006. Nutritional evaluation of tropical green seaweeds *Caulerpa lentillifera* and *Ulva reticulata*. *J. Nat. Sci.* 40: 75-83.
- Rensing, S.A. 2016. Why we need more non-seed plant models. *New Phytol.* 216: 355-60.
- Reynolds, C.S., Huszar, V., Kruk, C., Naselli-Flores, L., and Melo, S. 2002. Towards a functional classification of the freshwater phytoplankton. *J. Plankton Res.* 24(5): 417-28.
- Rincón-Díaz, N., Sanchez, J.V.M., Gavio, B., Chasqui, L. 2021. Diversity of benthic macroalgae in the Colombian Pacific: a study of a rocky reef flora. *Nova Hedwigia*. 112: 1-15.
- Roberts, D.A., Paul, N.A., Bird, M.I., and de Nys, R. 2015. Bioremediation for coal-fired power stations using macroalgae. *J. Environ. Manage.* 153: 25-32.

- Rodrigues, M. A., and da Silva Bon, E. P. 2011. Evaluation of *Chlorella* (Chlorophyta) as source of fermentable sugars via cell wall enzymatic hydrolysis. *Enzyme Res.* DOI: 4061/2011/405603.
- Rogers, M.B., Gilson, P.R., Su, V., McFadden, G.I., and Keeling, P.J. 2007. The complete chloroplast genome of the Chlorarachinophyte *Bigeloviella natans*: Evidence for independent origins of Chlorarachinophyte and Euglenid secondary endosymbionts. *Mol. Bio. Evol.* 24(1): 54-62.
- Rubin, B.E., Wetmore, K.M., Price, M.N., Diamond, S., Shultzaberger, R.K., Lowe, L.C., Curtin, G., Arkin, A.P., Deutschbauer, A., and Golden, S.S. 2015. The essential gene set of a photosynthetic organism. *P. Natl. A. Sci.* 112(48): E6634-43.
- Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A., Manhart, J.R. 2008. Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug *Elysia chlorotica*. *P. Natl. A. Sci.* 105(46): 17867-71.
- Rybak, A. 2018. Species of *Ulva* (Ulvophyceae, Chlorophyta) as indicators of salinity. *Ecological Indicators.* 85: 253-61.
- Rousseaux, C.S. and Gregg, W.W. 2014. Interannual variation in phytoplankton primary production at a global scale. *Remote Sens.* 6(1): 1-19.
- Sahoo, D., Elangbam, G., and Devi, S.S. 2012. Using algae for carbon dioxide capture and bio-fuel production to combat climate change. *Phykos.* 42(1): 32-8.
- Santelices, B. and Marquet, P. 2003. Seaweeds, latitudinal diversity patterns, and Rapoport's Rule. *Divers. Distrib.* 4(2): 71-5.

- Sathivel, A., Raghavendran, H.R.B., Srinivasan, P., and Devaki, T. 2008. Anti-oxidative and anti-hyperlipidemic nature of *Ulva lactuca* crude polysaccharide on d-Galactosamine induced hepatitis in rats. *Food Chem. Toxicol.* 46(10): 3262-7.
- Saunders, G.W. and Kucera, H. 2010. An evaluation of *rbcL*, *tufA*, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. *Cryptogamie, Algologie.* 4: 487-528.
- Schattner, P., Brooks, A.N., and Lowe, T.M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33 (S2): W686-9.
- Scragg, A.H., Morrison, J., and Shales, S.W. 2003. The use of a fuel containing *Chlorella vulgaris* in a diesel engine. *Enzyme Microb. Tech.* 33(7): 884-9.
- Sebe-Pedros, A., Degnan, B.M., and Ruiz-Trillo, I. 2017. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* 18: 498-512.
- Servais, T. and Harper, D.A.T. 2018. The Great Ordovician Biodiversification Event (GOBE): definition, concept, and duration. *Lethaia*: 51: 151-64.
- Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner K.E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G.J., Guan, Y., Shen, Y., Evgrafov, O.V., Knowles, J.A., Thibaud-Nissen, F., Schneider, V., Yu, C.Y., Zhou, L., Eichler, E.E., So, K.F., and Wang, K. 2016. Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* 7: 12065.
- Shuba, E.S., and Kifle, D. 2018. Microalgae to biofuels: ‘Promising’ alternative to renewable energy, review. *Renew. Sust. Energ. Rev.* 81(1): 743-55.

- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *J. Bioinform.* 31(19): 3210-2.
- Smetacek, V. and Zingone, A. 2013. Green and golden seaweed tides on the rise. *Nature*. 504: 84-8.
- Spoerner, M., Wichard, T., Bachhuber, T., Stratmann, J., and Oertel, W. 2012. Growth and thallus morphogenesis of *Ulva mutabilis* (Chlorophyta) depends on a combination of two bacterial species excreting regulatory factors. *J. Phycol.* 48: 1433-47.
- Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32 (S2): W309-12.
- Steinhagen, S., Barco, A., Wichard, T., and Weinberger, F. 2018. Conspecificity of the model organism *Ulva mutabilis* and *Ulva compressa* (Ulvophyceae, Chlorophyta). *J. Phycol.* 55(1): 25-36.
- Sterck, L., Billiau, K., Abeel, T., Rouzé, P., and de Peer, Y.V. 2012. ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods*. 9(11): 1041.
- Stratmann, J., Paputsoglu, G., and Oertel, W. 1996. Differentiation of *Ulva mutabilis* (Chlorophyta) gametangia and gamete release are controlled by extracellular inhibitors. *J. Phycol.* 32: 1009-21.
- Sun, L., Fang, L., Zhang, Z., Chang, X., Penny, D., and Zhong, B. 2016. Chloroplast Phylogenomic Inference of Green Algae Relationships. *Sci. Rep.* 6: 20528.

- Tait, K., Joint, I., Daykin, M., Milton, D.L., Williams, P., and Cámara, M. 2005. Disruption of quorum sensing in seawater abolishes attraction of zoospores of the green alga *Ulva* to bacterial biofilms. *Environ. Microbiol.* 7(2): 229-40.
- Tanaka, K., Ohno, M., and Largo, D.B. 2020. An update on the seaweed respires of Japan. *Bot. Mar.* 63(1): 105-17.
- Tan, I.H., Blomster, J., Hansen, G., Leskinen, E., Maggs, C.A., Mann, D.G., Sluiman, H.J., and Stanhope, M.J. 1999. Molecular phylogenetic evidence for a reversible morphogenetic switch controlling the gross morphology of two common genera of green seaweeds, *Ulva* and *Enteromorpha*. *Mol. Biol. Evol.* 16(8): 1011-18.
- Tanner, C. E. 1979. The taxonomy and morphological variation of distromatic ulvaceous algae (Chlorophyta) from the northeast Pacific. *University of British Columbia*.
<https://open.library.ubc.ca/collections/ubctheses/831/items/1.0094814>.
- Tanner, C.E. 1986. Investigations of the taxonomy and morphological variation of *Ulva* (Chlorophyta): *Ulva californica* Wille. *Phycologia*. 25(4): 510-20.
- Teichberg, M., Fox, S.E., Olsen, Y.S., Valiela, I., Martinettos, P., Iribarnes, O., Muto, E.Y., Petti, M.A.V., Corbisier, T.N., Soto-Jiménez, M., Páez-Osuna, F., Castro, P., Freitas, H., Zitelli, A., Cardinaletti, M., and Tagliapietra, D. 2010. Eutrophication and macroalgal blooms in temperate and tropical coastal waters: nutrient enrichment experiments with *Ulva* spp. *Glob. Change Biol.* 16: 2624-37.
- Titlyanov, E.A., Geebova, N.T., and Kotlyarova, L.S. 1975. Seasonal changes in structure of the thalli of *Ulva fenestrata*. *Ekologiya* 6: 320-24.

- Trachana, K., Larsson, T.A., Powell, S., Chen, W.H., Doerks, T., Muller, J., Bork, P. 2011. Orthology prediction methods: a quality assessment using curated protein families. *Bioessays*. 33: 769-80.
- Tragin, M. and Vaultot, D. 2018. Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. *Sci. Rep.* 8: 14020.
- Tran, L.A.T., Vieira, C., Steinhagen, S., Maggs, C.A., Hiraoka, M., Shimada, S., Nguyen, T.V., De Clerck, O., and Leliaert, F. 2022. An appraisal of *Ulva* (Ulvophyceae, Chlorophyta) taxonomy. *J. Appl. Phycol.* 34: 2689-703.
- Trivedi, N., Baghel, R.S., Bothwell, J., Gupta, V., Reddy, C.R.K., Lali, A.M., and Jha, B. 2016. An integrated process for the extraction of fuel and chemicals from marine macroalgal biomass. *Sci. Rep.* 6: 30728.
- Umen, J.G. 2014. Green algae and the origins of multicellularity in the plant kingdom. *CHS Perspect. Biol.* DOI: 10.1101/cshperspect.a016170.
- Valente, L.M.P., Gouveia, A., Rema, P., Matos, J., Gomes, E.F., and Pinto, I.S. 2006. Evaluation of three seaweeds *Gracilaria bursa-pastoris*, *Ulva rigida* and *Gracilaria cornea* as dietary ingredients in European sea bass (*Dicentrarchus labrax*) juveniles. *Aquacult.* 252(1) :85-91.
- Vesty, E.F., Kessler, R.W., Wichard, T., and Coates, J.C. 2015. Regulation of gametogenesis and zoosporogenesis in *Ulva linza* (Chlorophyta): comparison with *Ulva mutabilis* and potential for laboratory culture. *Front. Plant Sci.* 6(15). DOI: 10.3389.fpls.2015.00015.
- Vieira, H.H., Bagatini, I.L., Guinart, C.M., Vieira, A.A.H. 2016. *tufA* gene as molecular marker for freshwater Chlorophyceae. *Algae*. 31(2): 155-65.

- Vijayaraghavan, K., Jegan, J.R., Palanivelu, K., and Velan, M. 2004. Copper removal from aqueous solution by marine green algae *Ulva reticulata*. *Electron. J. Biotechnol.* 7(1): 61-71.
- Villares, R., Puente, X., and Carballeira, A. 1999. Nitrogen and phosphorus in *Ulva* sp. In the Galician Rias Bajas (northwest Spain): Seasonal fluctuations and influence on growth. *Bol. Inst. Esp. Oceanogr.* 15: 337-41.
- Vishwakarma, J., Parmar, V., Vavilala, S.L. 2019. Nitrate stress-induced bioactive sulfated polysaccharides from *Chlamydomonas reinhardtii*. *J. Biomed. Res.* 6: 7-16.
- Voigt, J., Liebich, I., Kieß, M., and Frank, R. 2001. Subcellular distribution of 14-3-3 proteins in the unicellular green alga *Chlamydomonas reinhardtii*. *Eur. J. Biochem.* 268: 6449-57.
- Voigt, J., Woestemeyer, J., and Frank, R. 2007. The chaotrope-soluble glycoprotein GP2 is a precursor of the insoluble glycoprotein framework of the *Chlamydomonas* cell wall. *J. Biol. Chem.* 282: 30381-92.
- Wang, B., Qiu, X.L., Peng, X., and Wang, F. 2018. Phytoplankton community structure and succession in karst cascade reservoirs, SW China. *Inland Waters.* 8(2): 229-38.
- Wang, S., Li, L., Li, H., Sahu, S.K., Wang, H., Xu, Y., Xian, W., Song, B., Liang, H., Cheng, S., Chang, Y., Song, Y., Cebi, Z., Wittek, S., Reder, T., Peterson, M., Yang, H., Wang, J., Melkonian, B., P, de Peer, Y.V., Xu, X., Wong, G.K.S., Melkonian, M., Li, H., and Liu, X. 2020. Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat. Plants.* 6: 95-106.

- Warren, R.L., Sutton, G.G., Jones, S.J.M. and Holt, R.A. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*. 23(4): 500-1.
- Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M., and Kriventseva, E.V. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res*. 41: D358-65.
- Weinberger, F., Steinhagen, S., Karez, R., and Bonthond, G. 2021. The *tufA* metabarcoding approach for *Ulva* and related seaweeds. ARPHA Conference Abstracts 4: e65495. <https://doi.org/10.3897/aca.4.e65495>.
- Weinshilboum, R. M. and Otterness, D. M. 1994. Sulfotransferase enzymes. Handbook of Experimental Pharmacology, ed. Kauffman (Berlin: Springer-Verlag). 45-78.
- Weiss, A., Costa, R. and Wichard, T. 2017. Morphogenesis of *Ulva mutabilis* (Chlorophyta) induced by *Maribacter* species (Bacteroidetes, Flavobacteriaceae). *Bot. Mar.* 60(2): 197-206.
- Wichard, T. and Oertel, W. 2010. Gametogenesis and gamete release of *Ulva mutabilis* and *Ulva lactuca* (Chlorophyta): regulatory effects and chemical characterization of the “swarming inhibitor”. *J. Phycol.* 46: 248-59.
- Wichard, T., Charrier, B., Mineur, F. Bothwell, J.H., De Clerck, O., and Coates. J.C. 2015. The green seaweed *Ulva*: a model system to study morphogenesis. *Front. Plant Sci.* 6(72).
- Wick, R.R., Shultz, M.B., Zobel, J. and Holt, K.E. 2015. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*. 31(20): 3350-2.

- Wickstead, B., Ersfeld, K., and Gull, K. 2003. The mitotic stability of the minichromosomes of *Trypanosoma brucei*. *Mol. Biochem. Parasit.* 132(2): 97-100.
- World Bank. 2013. Fish to 2030. Prospects for Fisheries and Aquaculture. World Bank Report No. 83177-GLB. Washington, DC. 80 pp.
- Wuebbles, D.J. and Jain, A.K. 2001. Concerns about climate change and the role of fossil fuel use. *Fuel Process. Technol.* 71: 99-119.
- Yamazaki, T., Ichihara, K., Suzuki, R., Oshima, K., Miyamura, S., Kuwano, K., Toyoda, A., Suzuki, Y., Sugano, S., Hattori, M. and Kawano S. 2017. Genomic structure and evolution of the mating type locus in the green seaweed *Ulva partita*. *Nature Sci. Rep.* 7: 11679.
- Ye, N., Zhang, X., Mao, Y., Liang, C., Xu, D., Zou, J., Zhuang, Z., and Wang, Q. 2011. ‘Green tides’ are overwhelming the coastline of our blue planet: taking the world’s largest example. *Ecol. Res.* 26: 477-85.
- Ye, N., Zhang, X., Miao, M., Fan, X., Zheng, Y., Xu, D., Wang, J., Zhou, L., Wang, D., Gao, Y., Wang, Y., Shi, W., Ji, P., Li, D., Guan, Z., Shao, C., Zhuang, Z., Gao, Z., Qi, J., and Zhao, F. 2015. *Saccharina* genomes provide novel insight into kelp biology. *Nat. Commun.* 6: 6986.
- Yu-Qing, T., Mahmood, K., Shehzadi, R., and Ashraf, M.F. 2016. *Ulva lactuca* and its polysaccharides: food and biomedical aspects. *J. Biol.* 6(1): 140-51.
- Xu, H., Miao, X., and Wu, Q. 2006. High quality biodiesel production from a microalga *Chlorella protothecoides* by heterotrophic growth in fermenters. *J. Biotechnol.* 126(4): 499-507.

- Zeroual, Y., Moutaouakkil, A., Dzairi, F.Z., Talbi, M., Chung, P.U., Lee, K., and Blaghen, M. 2003. Biosorption of mercury from aqueous solution by *Ulva lactuca* biomass. *Bioresour. Technol.* 90: 349-51.
- Zhang, X., Ye, N., Liang, C., Mou, S., Fan, X., Xu, J., Xu, D., and Zhuang, Z. 2012. *De novo* sequencing and analysis of the *Ulva linza* transcriptome to discover putative mechanisms associated with its successful colonization of coastal ecosystems. *BMC Genomics.* 13: 565-78.
- Zhang, J., Huo, Y., Wu, H., Yu, K., Kim, J.K., Yarish, C., Qin, Y., Liu, C., Xu, R., and He, P. 2014. The origins of the *Ulva* macroalgal blooms in the Yellow Sea in 2013. *Mar. Pollut. Bull.* 89: 276-83.
- Zodape, S.T. 2001. Seaweeds as a biofertilizer. *J. Sci. Ind. Res.* 60: 378-82.

DIGITAL APPENDIX - TABLE OF CONTENTS

The digital appendix is available for download upon request and is archived with Durham University.

Folder 1.0: Mpileup searchable databases

This folder contains files of a novel, searchable database of each replicate Langston or Ward sequence file (1 and 2) alongside the *U. mutabilis* sequence. The user may sort for nucleotides in which one, both, or neither replicate sequence matches the specific *U. mutabilis* base at a given position.

File 1.1 mpileup ward mutabilis.xlsx

File 1.2 mpileup langston mutabilis.xlsx

Folder 2.0: genome files

This folder contains fasta files for the two novel pseudoassemblies. All other genomes used in this thesis are publicly available (NCBI).

File 2.1 ward pseudoassembly.fa

File 2.2 langston pseudoassembly.fa

Folder 3.0: pairwise comparative genomics

This folder contains the PAF file for each possible pairwise comparison among genomes used in this study, along with a visual for a single pairwise comparison referenced earlier.

File 3.1	ward mutabilis PAF.tsv
File 3.2	langston mutabilis PAF.tsv
File 3.3	ward langston PAF.tsv
File 3.4	caulerpa mutabilis PAF.tsv
File 3.5	prolifera mutabilis PAF.tsv
File 3.6	prolifera mutabilis midline analysis.jpg
File 3.7	caulerpa prolifera PAF.tsv
File 3.8	langston prolifera PAF.tsv
File 3.9	ward prolifera PAF.tsv