Annotation of Marine Eukaryotic Genomes

By

Bram Verhelst

Department of Plant Biotechnology and Bioinformatics
University of Ghent

A dissertation submitted to the University of Ghent in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Sciences, specialisation Bioinformatics.

September 2015

Exam Committee

Prof. Dr. Peter Dawyndt (chair)
Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics (WE02), Ghent University, Belgium

Prof. Dr. Yves Van de Peer (promotor)
Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University, Belgium

Prof. Dr. Olivier De Clerck (secretary)
Faculty of Sciences, Department of Biology (WE11), Ghent University, Belgium

Prof. Dr. Pieter De Bleser
Faculty of Sciences, Department of Biomedical Molecular Biology (WE14), Ghent University, Belgium

Prof. Dr. Tim De Meyer
Faculty of Bioscience Engineering, Department of Mathematical Modelling, Statistics and Bioinformatics (BW10), Ghent University, Belgium

Dr. Gwenaël Piganeau
Université Pierre et Marie Curie (UPMC), UMR 7232, Observatoire Océanologique, Banyuls-sur-Mer, France

Dr. Pierre Rouzé
Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University, Belgium

Pro-forma proposal submitted to the Education Commission Biochemistry and Biotechnology (OCBB) at the 6th of September, 2015, and approved on the 8th of September, 2015.
The Faculty Council approved the proposal on the 23th of September, 2015.
The closed thesis defense was concluded at the 25th of November, 2015.
The public thesis defense is scheduled to take place at the 29th of January, 2016.

Publications

Publications taken up in this thesis

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M. F., Piganeau, G., Rouzé, P., Da Silva, C., Wincker, P., Van de Peer, Y. & Vandepoele, K. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13.** doi:10.1186/gb-2012-13-8-r74, R74 (2012) *(impact factor 10.288; journal ranking: 4/160 in biotechnology & applied microbiology)*

Verhelst, B., Van de Peer, Y. & Rouzé, P. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biology and Evolution* **5.** doi:10.1093/gbe/evt189, 2393–2401 (2013) *(impact factor 4.532; journal ranking: 11/46 in evolutionary biology)*

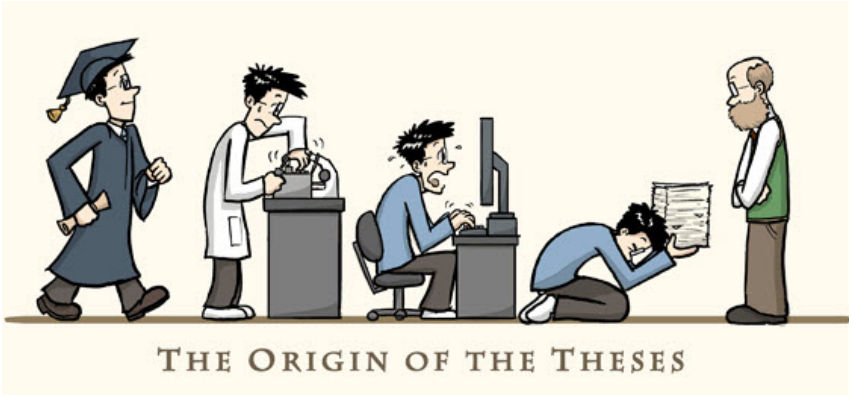
Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F., Carre, I., Chateau, A., Eyre-Walker, A., Grimsley, N., Moreau, H., Piegu, B., Rivals, E., Schackwitz, W., Van de Peer, Y. & Piganeau, G. An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* **15.** doi:10.1186/1471-2164-15-1103, 1103 (2014) *(impact factor 3.986; journal ranking: 26/162 in biotechnology & applied microbiology)*

Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y.-C., Bayer, T., Collen, J., Dattolo, E., De Paoli, E., Dittami, S., Maumus, F., Michel, G., Kersting, A., Lauritano, C., Lohaus, R., Töpel, M., Tonon, T., Vanneste, K., Amirebrahimi, M., Brakel, J., Boström, C., Chovatia, M., Grimwood, J., Jenkins, J. W., Jüterbock, A., Mraz, A., Stam, W. T., Tice, H., Bornberg-Bauer, E., Green, P. J., Pearson, G. A., Procaccini, G., Duarte, C. M., Schmutz, J., Reusch, T. B. H. & Van de Peer, Y. Genome re-engineering from land to sea by the seagrass Zostera marina. *Nature.* Accepted (2015) *(impact factor 41.456; journal ranking: 1/56 in multidisciplinary sciences)*

Publications not explicitly taken up in this thesis

Vandepoele, K., Van Bel, M., Richard, G., Van Landeghem, S., Verhelst, B., Moreau, H., Van de Peer, Y., Grimsley, N. & Piganeau, G. pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **13.** doi:10.1111/1462-2920.12174, 2147–2153 (2013) (impact factor 6.240; journal ranking: 13/119 in microbiology)

Acknowledgements



©Jorge Cham (PhDcomics)

The last ~~four~~ five years have been an amazing experience. The academic setting we're in provides a unique combination of highly-motivated eager individuals and limitless scientific possibilities, creating not only high-impact publications, but a truly great working environment. I have had the opportunity to work on data no-one else had before me, was involved in international consortia, and helped provide other scientists with the building blocks to initiate their own research. This in itself is very rewarding (stereotype, but true!). The yearly (scientific!) brainstorm meetings and (scientific!) trips abroad to Banyuls-sur-Mer and Edinburgh were an added bonus... In the next paragraphs, I would like to thank colleagues and family for their never-ending support that helped me keep moving forward.

I am highly aware of the fact that this section is usually written in a more format style, but introducing a bit of humour into an otherwise serious topic is a necessary evil. Therefore it seemed fitting to begin this section with a comic from PhDcomics, hilarious but – sadly enough – often very realistic (<sarcasm>which does not imply I agree with them!</sarcasm>). You might be interested to hear that I've spent almost as much time writing this section than I did the rest of the PhD (no more explicit sarcasm tags from now on!), because this is often the most-read section of the entire PhD thesis. To those reading this, I would like to apologise in advance – not really – for any of the inside-references mentioned in the next paragraphs: "you simply had to be there"! Oh, and to annoy any of the 'python rulez' folk I decided to introduce some Perl code... enjoy!

if($name =~/(University|N2N|VIB)/i) {

> I would like to thank Ghent University and the multi-disciplinary research partnership 'From nucleotides to networks' (N2N), now Bioinformatics Institute Ghent N2N, for providing the Dehousse scholarship that allowed me to start this PhD. I would also like to thank the Flemish Institute for Biotechnology (VIB) whose funds helped me span the initial months!

} elsif($name eq "Yves") {

Yves. When my application for an IWT grant was denied, you decided I was worth the investment. In hindsight this was a big risk as you could not know what my output would be. As my promotor we had many sciensy conversations, but shared many laughs as well. I was even allowed to touch the holiest of holies, your personal macbook :-) Thanks for allowing me to stay in your lab all these years, thanks for the support, and I really wish you all the best where-ever the South-African wind will take you!

} elsif($name eq "Pierre") {

Pierre, or, as I referred to you whenever I had to explain the lab hierarchy, ”my french boss”. Along the way, I really started to admire your keen eye, the level of dedication to science, and the many results you obtained, some of which are featured in this thesis’ chapters. You recently retired and I wish you, your family and your horses all the best. You have provided science a great service!

} elsif(@exam_committee ~~ $name) {

I sincerely thank each member of the exam committee for agreeing to be part of this final chapter of my PhD. Thank you for reading the manuscript and challenging me with some mind-boggling scientific questions.

} elsif(@beg ~~ $name) {

The Bioinformatics and Evolutionary Genomics group, or just ’BEG’. When I entered the lab, the group was gigantic. Over the years, many post-docs have left – no causal relation with myself! – leaving behind a small gathering of nerdy but friendly folk! I’ve enjoyed my days in the lab and would like to thank several lab members!

Let’s start of with the annotation group. When I joined the lab as a master student, I was introduced to Pierre, Yao-Cheng, Lieven and Stephane. Throughout all these years, these "senior members" have been the go-to persons when I was in dire need for scientific advice, or just for a good conversation. For some reason, the open spot on your island was always taken the first months I was there, so I had to go sit elsewhere. Know that I always longed for that spot! <sarcasm tags laying around here somewhere ;-)> No seriously, you guys were great colleagues!

Next up, the Fellowship of the Lounge. Lieven, Ken, Jan, Jonas, Bo, Michiel, Thomas, Nicolas (†), Klaas, Ronnie, Bram^2, Frederik,... the daily 15h lounge moments were a true testament to the human mind and the depths it ventures in to. I would like to go into detail here, but I’m afraid certain quotes are just inappropriate for a thesis (or for that matter, any verbal & non-verbal medium!), especially when taken out of context <r/evenwithcontext>. To this day, I still don’t know why you presented me with that great totem-name "Ranzige Das".

Finally I would like to wish Phuong and Bing, my two former island colleagues, all the best! And to everyone in the lab, remember me whenever you are standing by the fridge and see ceiling-cat looking down upon you while thinking "Damn, who is going to fill the fridge from now on?" Or "These plants look a bit dead-ish. Isn’t anyone watering them?" Be warned, as "Minister of Plants" I will be electing a successor soon! You could be next!

} elsif(@it ~~ $name) {

"Have you tried turning it off and on again?" Thanks guys for keeping everything running (well, most of the time :p Thank god we had that power generator!). Even after five years, I still get confused by the intricate web of shares, mounts and servers. It seems to have its own will,

and IT is there to tame it (again, most of the time...). Special thanks to Frederik, who put up with my many many many questions and demands for new software installations, and who introduced me to D. I'm sure the helpdesk INBOX will be much lighter from now on!

} elsif(@collaborators ~~ $name) {

A special thank you to the people involved in all the genome projects, both wetlab and drylab scientists. It's been an absolute treat to work with you. I hope I wasn't too computer-geeky for you when trying to explain what I did or how to proceed on ORCAE. A special thanks to Banyuls-sur-Mer (Gwenael, Hervé) and to the Edinburgh SynthSys people (Matthew, Thierry, Andrew) for the scientific visits!

} elsif($name eq "Ken") {

Ken, my partner-in-crime. I joined your year group after my disastrous first year at university, but didn't notice you in the beginning <queue height joke here>. We started playing cards during breaks together with Jonas and Thomas. Eventually we both choose bioinformatics as our major in the master years at university, which resulted in classes where it would often be just the two of us. Professors told us we were like a couple, constantly bickering, and later on, during our PhD life, I might have even said I was your 'work-wife'. We both continued our academic career in the same lab environment, but on different topics. Recently we got divorced – I'm continuing a theme here! – and you moved out to Bayer while I stayed behind. We still try to get together each week to play badminton because you aspire to the same level of body perfection that I do <queue fat joke here to even out the previous height joke ;-)> That reminds me, it's been a terrible long time since we had one of our movie days, and Ken... the university won't change the way they calculate the points on the diploma ;-)

} elsif(@family ~~ $name) {

My family and the running gag. A running gag is a joke that is only mildly funny at first, but becomes funnier each time it is repeated. And boy, did it get repeated... Several gag themes circulate such as "you should be thanking us for paying taxes so you can continue your research (so when does this money start flowing back to us?)", and "Bram doesn't really work, he sits behind a computer all day and browses the internet". I enjoyed such banter because it allowed me to reciprocate the teasing and you all were genuinely interested in what I did ;-) DAAROM DUS, BEDANKT VOOR DE STEUN GEDURENDE AL DIE JAREN!. I sincerely hope you did not bring the giant wooden doll – a family tradition reserved for anniversaries – to my PhD ceremony, or I will 'sink through the ground from embarrassment'.

} elsif(@parents ~~ $name) {

Moe & Va, if you hadn't talked some sense in to me after my first year at university I probably wouldn't be standing here today. I got my act together and decided to go for a PhD after graduating university. For four years I lived mainly in Gent but returned each weekend to 'Hotel Mama', all comforts included. I know I haven't been the easiest to live with, especially since I had to be driven everywhere (who needs a driving license?) and my biorhythm seemed to be the complete opposite of yours (you must have thought I was a vampire at one point). Do know that I appreciate everything you have done for me during all these years! BEDANKT VOOR ALLES!

} elsif($name eq "Ben") {

Ben, you get your own paragraph! We always got along really well. When we were young, you pushed me around on a cart (I was better in delegating!) and now, I carry you around

whenever you jump on my back. Although younger, you moved out quicker than I did, getting your own house and recently a baby boy. I really wish you, Jolien and Stan the very best of futures.

} elsif($name eq "Joke") {

Joke, I know I have been absent many weekends and evenings the past month(s) locked behind my computer, writing this thesis. If only they could give me my diploma without it (is this an option?), then I wouldn't have to leave your side so often. We're been living together for almost a year now and it's really great coming home to you each day! IK ZIE JE GRAAG, LIEFJE! BEDANKT VOOR DE STEUN.. Also, thanks for the language advice (I told you I was going to mention it!).

} elsif($name eq "BLAST") {

I would like to thank you, BLAST, for the many great scripts and pipelines we built together! I confess I cheated on you with your faster, be it more buggy, sibling Decypher, but know there is no application you can not be used for!

} elsif($name =~/LaTeX /i) {

I would like to thank the LaTeX community for providing the tools essential to the creation of this document. I couldn't imagine doing this in Word.

} else {

I am sorry I forgot to mention you by name. I am sure your support is what kept me going in difficult times!

}

I love doing what I do now, assembling and annotating, everyday another challenge. If I could sign up for the rest of my life, I would do so in a heartbeat. With these words, this section has come to and end. Let's move on to the more serious matters!

Summary

Genome annotation includes two essential steps: finding the genes on the DNA (structural annotation = gene prediction) and adding functional information to the genes (functional annotation). Gene prediction is a substantial cornerstone of each genome project. Custom annotations are produced using a combination of training sets, *ab initio* gene finders, various extrinsic sources, combination algorithms and manual curation. Afterwards, genomes are probed and analysed to discover the true nature of the organism. Within this thesis, we describe the detailed genome analysis of seven organisms, together with the unique adaptations that helped shape their evolution.

The unicellular micro-algae *Ostreococcus* (*tauri*, *lucimarinus*, RCC899), *Micromonas* (CCMP1545, RCC299) and *Bathycoccus* (*prasinos*) represent the prasinophytes, one of the earliest branches in the green tree of life. They exhibit reduced cellular complexity (single chloroplast, mitochondrium, Golgi body, ...) and reduced genome sizes (from 12 to 22 Mb) coupled to a high coding density. The genomes of these Mamiellales species have been sequenced and annotated, where after two peculiar regions were noted that stood out in comparison to the rest of the genome: the outlier chromosomes. They exhibit heterogeneity in terms of sequence content, gene expression, gene structure, gene origin, and are even known to change in chromosome length. The Big Outlier Chromosome, or BOC, is a low GC% region that contains highly expressed genes that are of vital importance to the organism. BOC genes contain more exons, and the introns are grouped into a class distinct from all other introns in the genome. Unlike the name suggests, BOC does not encompass the entire chromosome: it contains the outlier region (BOC1) flanked on one or either side by a region with non-outlier characteristics (BOC0). The reduced level of recombination (~low-GC%) and extreme gene shuffling are reminiscent of a sex chromosome or species barrier. The latter would imply that correct pairing of two BOC chromosomes in a new prasinophyte offspring is only possible if the two gametes originate from the same parent line. Another outlier chromosome is most often the smallest chromosome within the genome, and is appropriately called Small Outlier Chromosome (SOC). Genes are often truncated, without known homologs and often linked to horizontal gene transfer. They function mostly in the modification of the cell surface (glycosylation, transmembrane proteins) and seem to play an important role in virus interaction. The concentration of such specific features into two outlier chromosomes seems to be unique for Mamiellales species.

Each Mamiellales species has two groups of introns. The Big Outlier Chromosomes contains a group of very small AT-rich introns, dubbed BOC1 introns, that are very distinct from canonical spliceosomal introns found in the rest of the genome. SOC usually does not feature introns. On top of this, the intron landscape in *Micromonas* is even more complex, with the addition of repeat-like introns that share many characteristics with transposable elements: the Introner Elements (IEs). They make up more than 50% of all CCMP1545 introns and are responsible for the 1 Mb surplus in genome size in relation to RCC299. Introner-Like Elements are found in fungi, and just as in *Micromonas*, different clades contain different families of IEs. The proposed propagation mechanisms are reminiscent of group II introns, with intron transposition on the mRNA level, and spliceosomal retrohoming on the DNA level. Metagenomic data revealed Presence/Absence Polymorphisms, a clear sign these repeat

introns are able to move about the genome and augment their numbers in the process. Over time, individual introns degenerate and loose their repeat-like nature, becoming indistinguishable from canonical spliceosomal introns.

Unlike the unicellular microalgae, the genome of the multicellular seagrass *Zostera marina* displays islands of genes separated from each other by large stretches of repeat elements: roughly 60% of the genome is composed of transposable elements. Various adaptations have enabled this seagrass to survive in marine environments. It has exineless pollen, an algal-like cell wall, and salt-tolerant ATPase antiporters. It has also lost many angiosperm features such as volatile production (terpenoids, ethylene), plant defense genes, UV damage repair enzymes, and stomatae.

Why these organisms? They are found globally and whether unicellular or multicellular, small genome or big genome, they are dominant members of their respective ecosystems. Knowledge on their inner workings will allow us to better understand not only these organisms, but also their partners within the ecosystem. Such data provides scientists with the necessary tools to develop much needed methods for bio-remediation, combating climate change, the prevention of erosion, the fight against harmful algal blooms, and many more. Furthermore, the unique features of each organism pose specific problems for gene prediction and require additional attention in order to produce a quality annotation.

Samenvatting

Genoom annotatie omvat twee essentiële stappen: het lokaliseren van genen op het DNA (gen predictie) en het toevoegen van functionele informatie (wat doet het gen?) aan de genen. Annotaties worden geproduceerd via een combinatie van *ab initio* gen-lokalisatie software, extrinsieke data, combinatorische algoritmes en manuële curatie. Nadien worden de genomen geanalyseerd om tot de ware aard van het organisme te komen. In deze thesis beschrijven wij de gedetailleerde genoom analyse van zeven organismen, alsook hun unieke aanpassingen die mee hun evolutie hebben bepaald.

De ééncellige micro-algen *Ostreococcus* (*tauri*, *lucimarinus*, RCC809), *Micromonas* (CCMP1545, RCC299) en *Bathycoccus* (*prasinos*) vertegenwoordigen de prasinofieten, een tak die zich aan de basis bevindt van al het groen leven. Hun cellulaire complexiteit is enorm gereduceerd (één chloroplast, mitochondrium, Golgi Apparaat) en hun genomen zijn erg klein (12 tot 22 Mb) maar met een erg hoge coderings-dichtheid. De genomen van deze Mamiellales soorten zijn gesequeneerd en geannoteerd, waarna twee uitzonderlijke regio's werden gedetecteerd die enorm verschillen van de rest van het genoom: de outlier chromosomen. Zij vertonen enorme verschillen op het vlak van gen expressie, gen structuur, gen oorsprong, en kunnen zelfs variëren in lengte. Het Grote Outlier Chromosoom, of BOC, is een genomische regio met laag GC gehalte waarop vele genen liggen die hoog geëxpresseerd worden en die van vitaal belang zijn voor het organisme. In tegenstelling tot wat de naam doet vermoeden omvat BOC echter geen volledig chromosoom: het omvat de outlier regio (BOC1) omringd aan één of beide zijden door een regio met niet-outlier eigenschappen (BOC0). De gereduceerde recombinatie (~laag GC gehalte) en extreme gen-herschikking zijn erg karakteristiek voor een sex chromosoom of species-barrière. Deze laatste stelt dat het correct paren van beide BOC chromosomen in een nieuwe prasinofiet nakomeling enkel mogelijk is indien beide gameten afkomstig zijn van dezelfde ouderlijke lijn. Een ander outlier chromosoom is vaak het kleinste chromosoom in het genoom, en wordt dan ook gelabeld als Klein Outlier Chromosoom (SOC). Genen zijn vaak afgekort, zonder gekende homologen en vaak gelinkt een horizontale gen transfer. Ze functioneren voornamelijk in modificatie van het cel oppervlak (glycosylatie, transmembranaire proteïnes) en spelen vaak een belangrijke rol in virus interacties. De concentratie van zulke specifieke eigenschappen in twee outlier chromosomen blijkt uniek te zijn voor Mamiellales soorten.

Elke Mamiellales soort heeft twee groepen van introns. Het Grote Outlier Chromosoom bevat een groep van zeer kleine AT-rijke introns, BOC1 introns genaamd, die erg verschillend zijn van de canonieke spliceosomale introns die je in de rest van het genoom terug vindt. SOC heeft meestal geen introns. Bovenop deze twee groepen bevat *Micromonas* een meer complex intron landschap, met de toevoeging van repeat-achtige introns die veel gelijkenissen vertonen met transposable elements: de Introner Elementen (IE). Vijftig percent van alle intronen in CCMP1545 zijn IEs, en zij zorgen voor een genoom-grootte die 1Mb groter is dan RCC299. Introner-achtige Elementen worden ook teruggevonden in schimmels, en juist zoals in *Micromonas* bevatten verschillende clades verschillende IE families. De vooropgestelde mechanismen waarmee deze IEs zich vermenigvuldigen doen ons sterk denken aan group II introns, met intron transpositie op het mRNA niveau, en spliceosomale retrohoming op het DNA niveau. Metagenomische data vertonen aanwezig/afwezig-polymorfismen, een duidelijk teken dat deze repeat introns zich doorheen het genoom kunnen

bewegen en zo hun aantal doen toenemen. In de loop der tijd verliezen deze introns hun repeat karakter en worden zo niet te onderscheiden van canonieke spliceosomale introns.

In tegenstelling tot de ééncellige microalgen bevat het genoom van het meercellige zeegrass *Zostera marina* eilandjes van genen, gescheiden van elkaar door lange regios van repeat elementen: ongeveer 60% van het genoom is opgemaakt uit transposable elements. Verscheidene aanpassingen laten dit zeegras toe om te overleven in een mariene omgeving. Het heeft exin-loos pollen, een alg-achtige celwand, en zout-tolerante ATPase antiporters. Het heeft ook vele eigenschappen verloren die eigen zijn aan Angiospermen, zoals de productie van vluchtige moleculen (terpenoïden, ethyleen), plant-verdedigings-genen, UV-schade herstel-enzymen en stomata.

Waarom juist deze organismen onderzoeken? Deze zijn globaal verspreid en of ze nu ééncellig of meercellig zijn, klein genoom of groot genoom, zij zijn de dominante soort in hun ecosysteem. De kennis over hoe zij functioneren, laat ons toe om niet enkel deze organismen beter te leren kennen, maar ook hun partners in het ecosysteem. Die data laat wetenschappers toe om methoden te ontwikkelen die nuttig zijn voor bio-remediatie, de strijd tegen de klimaatverandering, het voorkomen van erosie, het begrijpen en bestrijden van harmful algal blooms, en zoveel meer. De unieke eigenschappen van deze organismen vereisen ook extra aandacht tijdens de gen predictie indien we een kwalitatieve annotatie willen afleveren.

Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

Abbreviations

AAAD	aromatic acid decarboxylase. 108
BAC	Bacterial Artificial Chromosome. 23
BOC	Big Outlier Chromosome. 44–47, 49, 50, 53, 54, 56–60, 67, 68, 70–73, 76, 77, 83, 132–136
CA	carbonic anhydrase. 103, 109
CAT	catalase. 109
CDK	Cyclin-dependent Kinase. 25
cDNA	complementary DNA. 33–35, 72, 75, 136
CDS	coding sequence. 14, 24, 68, 69, 92–95, 98
CE8	carbohydrate esterase 8. 109
CM	co-variance model. 15
DBG	de Bruijn graph. 9, 10, 17, 97
dsDNA	double-strand DNA. 32, 50, 136
EC	Enzyme Commission. 114
EIN3	Ethylene Insensitive 3. 108
EST	Expressed Sequence Tag. 3, 12, 14, 15, 19, 34, 42, 45, 47, 53–56, 60, 67, 69, 70, 76, 99, 113, 114, 130, 135
FPKM	Fragments Per Kilobase of transcript per Million fragments mapped. 112, 113, 118, 119
GFF	General Feature Format. 14, 131
GHMM	General Hidden Markov Model. 14
GO	Gene Ontology. 17, 44, 48, 53, 56, 77, 113, 114
HGT	horizontal gene transfer. 18, 39, 47–50, 52, 54, 133
HMM	Hidden Markov Model. 13–15, 76
HMT	half-metallothionein. 121
ICM	Interpolated Context Model. 13
IE	Intron Element. 27, 65, 67, 69–72, 74–77, 83, 84, 130, 135–138
IEP	intron-encoded protein. 33, 69, 70, 137
ILE	Intron-like Element. 74, 137, 138
IMM	Interpolated Markov Model. 13, 14, 113
InE	Introns-Early. 29–31, 73
InL	Introns-Late. 30, 31, 73, 74
IR	inverted repeat. 122
ITS	Internal Transcribed Spacer. 42
JGI	Joint Genome Institute. 41, 44, 76, 97, 110, 111, 130
KDE	kernel density estimate. 107, 115
LEA	late embryogenesis abundant. 109
LHC	Light-harvesting complex. 25, 109
lncRNA	long non-coding RNA. 15

LSC long single copy region. 122
LUCA Last Universal Common Ancestor. 29–31
MCMC Markov Chain Monte Carlo. 115
miRNA micro-RNA. 15, 32, 106, 113
MM Markov Model. 12–14, 113
MMC Meristemoid Mother Cell. 108
mRNA messenger RNA. 15, 32, 34, 53, 72, 136
MSA multiple sequence alignment. 9, 15, 54, 55, 76
MT metallothionein. 109, 110, 121
NBS-LRR nucleotide binding site-leucine rich repeat. 103, 108
NCBI National Center for Biotechnology Information. 48, 53, 54, 71, 76, 99, 103, 112
ncRNA non-coding RNA. 15, 32, 130
NGS Next-Generation Sequencing. 9, 11, 12, 18, 89, 129, 133
NII Nearly Identical Intron. 74
NMD nonsense-mediated mRNA decay. 31, 34, 70, 74
NPQ non-photochemical quenching. 109
OLC overlap-layout-consensus. 9–11, 17
ORCAE Online Resource for Community Annotation of Eukaryotes. 15, 17, 39, 65, 76, 89, 95, 99, 103, 114, 115
ORF open reading frame. 12, 23, 29, 32, 48, 69, 96
PacBio Pacific Biosciences. 10, 11, 129, 135
PAP Presence/Absence Polymorphism. 65, 71, 72, 77, 138
PFGE Pulsed-Field Gel Electrophoresis. 18, 23, 42, 43
pre-mRNA precursor messenger RNA. 34, 67, 74, 75, 135
PWM Position Weight Matrix. 13
rDNA ribosomal DNA. 23, 29, 41, 111, 133, 137, 138
RISC RNA-induced Silencing Complex. 15
RNP ribonucleoprotein. 33
rRNA ribosomal RNA. 15, 21, 23, 24, 56, 75, 94
RSI regular spliceosomal intron. 34, 136, 137
RT Reverse Transcriptase. 29, 33, 34, 136, 137
SMRT Single-molecule real time. 10, 11, 129, 135
snRNA small nuclear RNA. 15, 31, 70
SOC Small Outlier Chromosome. 18, 44, 45, 47, 50, 54, 55, 59, 60, 70, 132, 133, 135
SSC short single copy region. 122
ssDNA single-strand DNA. 136
SuSy sucrose synthase. 109
SUT sucrose transport. 109
SVM Support Vector Machine. 13
TE transposable element. 12, 23, 27, 42, 72, 74, 75, 95, 98, 106, 112, 113, 117, 119, 120, 129, 137
tRNA transfer-RNA. 15, 29, 56, 95, 114, 122
UTR untranslated region. 12, 69, 99, 114
UV ultra-violet. 25, 103, 109
UVR ultra-violet resistance. 103, 109
WAM Weight Array Matrix. 13, 113
WGD whole genome duplication. 106, 107, 114, 115
WGS whole-genome shotgun. 9, 42, 106
WWAM Windowed WAM. 13

CHAPTER 1

AIMS & THESIS OUTLINE

Marine eukaryotes display a wide range of morphologies and lifestyles (e.g. unicellular vs. multicellular, mobile vs. stationary). Equally interesting, the genomes that encode for these properties also display a wide variety of characteristics. The goal of this PhD was to annotate the genomes and document the gene repertoire and genome sequence features of several marine eukaryotes, both unicellular and multicellular, in order to locate unique properties or adaptations that could provide an insight into how these organisms evolved and adapted to their environment.

Each genome project starts of with an assembly of the genome sequence of interest. With the help of different read libraries, genomes are carefully reconstructed. With the help of extrinsic data (Expressed Sequence Tags (ESTs), proteins, genomes of other organisms), models are trained and gene structures will be predicted. Functional annotation will try to add biological meaning to the gene structures (e.g. their function or place in a pathway). The gene models and functions resulting from the annotation process already provide a big resource for scientists that wish to delve deeper into the biology of the organism.

Once a compendium of genes has been constructed, we can mine the data to find 'interesting' (and hopefully unique) properties such as genes/pathways that have been lost or gained (~gene family analysis), aberrant gene structures (e.g. disproportionate amount of single-exon genes, huge introns, no intergenics,...) and genomic heterogeneity (i.e. parts of the genome display vastly different characteristics compared to the rest). Subsequently, these properties are compared to other closely-related organisms to determine the evolution (i.e. when did these properties originate? how did they evolve?), functionality (i.e. why can we find them in the genome?) and uniqueness (i.e. to which degree are the properties shared amongst the closely-related organisms?). Ultimately, the data allows us to draw conclusions on how these properties could have benefited the organism of interest and help to shape its evolution.

In **chapter 2**, information on the assembly and annotation process (structural and functional) is provided, as well as a basic description of the marine eukaryotes that are under investigation. Because introns are an intricate part to several of the chapters, a small introduction on introns and splicing has also been added.

Chapter 3 focuses on the genome of *Bathycoccus prasinos* and the presence of outlier chromosomes. These outlier chromosomes are characterised in detail and compared with other Mamiellales species

(*Micromonas pusilla*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*).

In **chapter 4** we focus on the genomes of *Micromonas pusilla* CCMP1545 and *Micromonas* sp. RCC299. The presence of several classes of repeat introns is documented, and their dispersal within other species and metagenomes is examined.

Another Mamiellales species, *Ostreococcus tauri*, is analysed in **chapter 5**. A novel genome assembly has been constructed, as well as a new gene annotation, providing a much needed update for the *Ostreococcus* scientific community.

Chapter 6 describes the assembly, annotation and unique adaptations of the seagrass *Zostera marina*.

Finally, the main conclusions are summarised and discussed in **chapter 7**. Intriguing research questions and future research perspectives are also provided.

CHAPTER 2

INTRODUCTION

Okay, then let's begin.
And here we go. And watch my hand. And one, two, three.

- Stewie Griffin

This chapter provides a basic overview of methods and tools used in assembly, gene prediction (both structural and functional annotation), as well as a brief biological description of the marine eukaryotes appearing in the thesis chapters. Introns are an intricate part of this thesis, which has prompted a short introduction into this subject matter at the end of this chapter.

2.1 The Annotation Process

The eukaryotic genome annotation project includes many steps [6], which will be explained in detail within the next sections, together with some best-practice approaches and often-used software. The procedures explained in this section have been used in all the genome projects that appear later on in this thesis.

2.1.1 Assembling the genome before annotation

Most genome projects use a whole-genome shotgun (WGS) sequencing approach to break the DNA up into numerous random small pieces. These pieces are sequenced to obtain different types of reads: single-end reads (no inner-mate distance), paired-end reads (short/medium inner-mate distance) and mate-pair reads (long inner-mate distance; 'jump' library). These reads are used in different stages of the assembly process i.e. the process of reconstructing the original genome sequence using powerful computer algorithms. Sufficient overlap between the reads is required in order to obtain a high-quality assembly. This implies that longer reads and higher coverage are always the best choice for genome projects, if the project budget allows it.

Read cleaning & merging

The latest chemistries of Next-Generation Sequencing technologies (Illumina HiSeq, Illumina MiSeq) produce reads with very low error rates (<1%). Nevertheless it is still necessary to perform quality control because non-removal of remnant PCR products (adapters, PhiX) and contaminants can have dire results for the final assemblies and influence downstream analyses [7]. Common practice includes clipping (adapter removal), trimming (removing read bases), quality trimming (removing read bases based on quality scores) and the subsequent filtering (are the reads still long enough after the former operations have been performed?). Tools such as ngsShoRT [8], bbduk [9], Trimmomatic [10] or fastq_quality_trimmer [11] can perform such operations with a varying degree of success [12]. Extensive quality reports can be generated using FastQC [13] or the NGS QC Toolkit [14].

The methods described in the previous paragraph will trim a read sequence based on a fixed length (trimming) or quality scores (quality trimming), reducing the amount of errors in the sequence. Such schemes still leave many single-base errors in the reads and needlessly discard valid sequence. Error-correction software such as QuorUM [15] or Blue [16] will utilise k-mer algorithms (e.g. suffix trees/arrays, k-mer-seed MSA) to refine single-base errors and reduce the overall read error rate [17, 18].

Finally, if the total sequenced fragment is shorter than the summed length of both read mates, the read mates overlap, which can cause issues within the assembly graph. It is best to merge both reads into a single, larger, read (FLASH [19], PANDAseq [20], COPE [21] and PEAR [22]).

De novo contig assembly

The overlap between the single-end and pair-end reads allows the assembly software to create continuous stretches of sequence, or contigs. The presence of repeat sequences within the genome – resulting in highly-similar reads – will confuse the assembly software: instead of re-generating the complete genome as several long contiguous sequences, a lot of smaller contigs will be produced.

There are two main categories of assembly algorithms [23]: overlap-layout-consensus (OLC) and de Bruijn graphs (DBGs). In OLC, the assembler identifies all reads that sufficiently overlap each other and organises this information into a graph (node = read; edge = overlap). Popular OLC assemblers are Celera Assembler [24] and Arachne [25], and were mostly used with SANGER-data. While quite old, the Celera assembler is continuously

updated and used in varying roles: CABOG (454 data [26]) and PBcR (PacBio/Nanopore [27]). A modern OLC assembler is SAGE [28].

DBG assemblers first chop up the reads into substrings (k-mers) which are turned into a graph much like the OLC approach (node = k-mer; edge = overlap by k-1). Popular DBG assemblers are ABySS [29], Velvet [30], SOAPdenovo2 [31], ALLPATHS-LG [32], CLC Assembly Cell (`http://www.clcbio.com/products/clc-assembly-cell/`) and DiscovarDeNovo (`http://www.broadinstitute.org/software/discovar/blog/`). The most important benefit of the DBG algorithm is the speed, though this comes with large caveats such as the loss of read coherence (some paths through the de Bruijn graph are inconsistent with respect to the input reads) and a very short overlap length k (dealing with repeats is more difficult). Some assemblers try to integrate both categories of algorithms, obtaining the computational efficiency of DBG and the flexibility of OLC [33].

Scaffolding

Mate-pair reads are derived from libraries with long fragment lengths. When both mate-pair reads map on different contigs, the contigs can be joined together into a scaffold if there is enough evidence. The area in between both contig segments of the scaffold – the length depends on the library fragment size – remains unknown and is constructed out of 'N' characters. Afterwards, gap filling/closing methods are often employed to resolve these gaps, either by recursively mapping reads onto both ends of the gap thereby slowly replacing the N characters into informative nucleotides (A, C, G or T) (e.g. GapFiller [34], Sealer [35]), or by using very long reads (offered by Pacific Biosciences (PacBio) and the MinION system). Scaffolding approaches are implemented in many assembly software, but standalone programs do exist (SSPACE [36], BESST [37], WiseScaffolder [38]). A recent evaluation concluded that these tools, whether standalone or integrated, would always fail to identify ~10% of all 'true' joins due to errors in read mapping, contig assembly errors and the implementation of the scaffolding algorithm itself [39].

Meta-assemblies, super-scaffolds and chromosomes

Results from different assembly strategies can be combined to form consensus meta-assemblies. Tools such as GARM [40], Slicesembler [41] and Metassembler [42] allow scientists to harness the strengths of individual assemblers and overcome algorithmic weak-points, thereby creating longer scaffolds. Due to the vast amount of time and resources required to run multiple assemblies and perform an all-vs-all comparison, the meta-assembly approach is often forsaken.

A high-density linkage map can anchor and correctly orient the scaffolds into super-scaffolds or pseudo-chromosome-like structures [43]. Gap filling/closing methods can again be applied to produce the final reference draft. Additionally, chromatin interaction data can also be used to scaffold draft assemblies [44].

A genetic linkage map requires a mapping population to generate recombination and genetic differences between related individuals. These individuals are genotyped at different markers where after linkage analysis can order these markers into a linear sequence (~chromosome) with the distances measured in centimorgans. A physical map will cut the sequence at specific sites using restriction enzymes or physical shearing e.g. sonication. The resulting fragments are separated, sized and finally pieced together to produce a physical map of the genome containing restriction site markers. A fine example of a physical map technology is 'Optical mapping' (BioNano Genomics) [45], which will 1) stretch and hold in place individual DNA molecules, followed by 2) a restriction enzyme digest, and 3) staining of the ordered segments. The fluorescence intensity of each segment is correlated to its size, creating an optical map of single DNA molecules. Several of such optical maps can be aligned to produce a genomic-size optical map *(Figure 2.1)*. This technology can significantly improve assemblies (MISSEQUEL [46]), and, in combination with long reads, produce near-complete finished genome assemblies [47, 48].

The rise of the long reads

Single-molecule real time (SMRT) sequencing technology has been around for several years [50]. Recently, both the technology has matured (PacBio RS II; P6-C4 chemistry; ~86% accuracy; 0.5–1Gb of data; average read

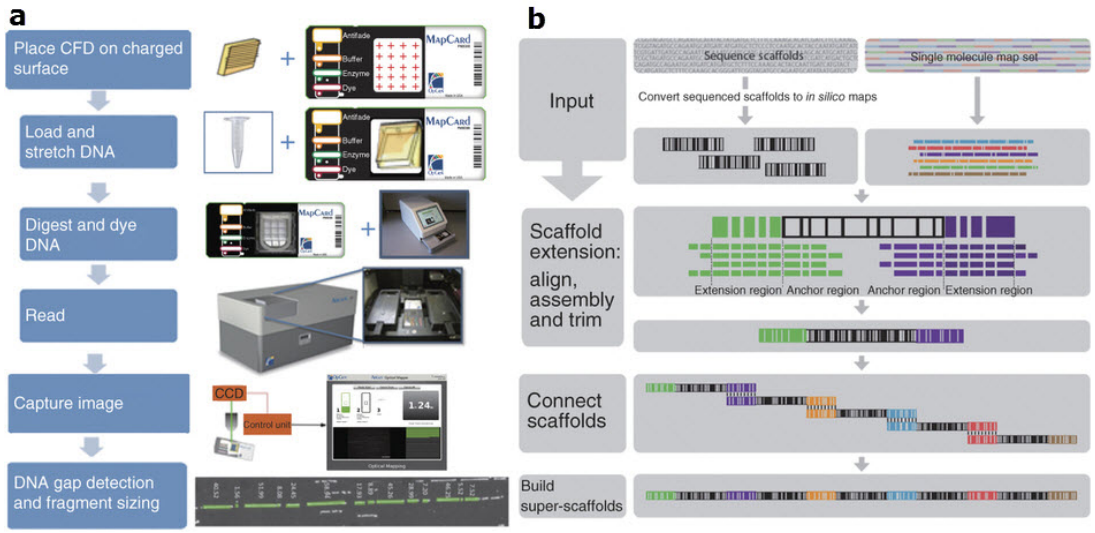


Figure 2.1. Whole-genome optical mapping. (a) Samples are loaded into a channel-forming device (CFD). Buffer fluid stretches and immobilises high molecular weight DNA. After digestion and staining, the digestion patterns are converted into single-molecule restriction maps. **(b)** *In silico* digested scaffolds are matched to the optical fragments allowing extension of the scaffolds into super-scaffolds. Source: [49]

lengths: 10–15kb), as well as the algorithms and software to cope with this kind of 'third generation sequencing' data. Many genome projects are now using a hybrid design, integrating PacBio reads together with the more traditional Illumina approach. While long reads are undoubtedly the future of assembly, some challenges still slow down its progress. Most PacBio algorithms use an OLC approach (see previous section) to arrange the long reads. The computational requirements to handle such data are huge, both in time and memory.

Depending on the coverage of long reads, methods are divided into four main categories [51–54]. *De novo* PacBio-only assembly is only possible with high coverage (>50x) which allows the algorithm to perform self-correction (shorter PacBio reads are used to clean longer PacBio reads) and finally assemble the cleaned reads. This category includes assemblers such as HGAP [55], PBcR/MHAP [27], Celera Assembler [24], Falcon, Dazzler and Sprai. A second category involves the hybrid approach, either using short high-fidelity Next-Generation Sequencing (NGS) reads (or pre-assembled contigs) to reduce the error rates in the long SMRT reads (LordEC [56], ECTools [57] and proovread [58]), or using long reads to scaffold draft NGS assemblies (DBG2OLC [59], SPADES [60] and Cerulean [61]). When the coverage is low, there are only two options: scaffolding (SSPACE-LongRead [62] and AHA [63]) or gap filling (PBJelly [64], GMcloser [65]).

Another platform offering long read sequencing is Ofxord Nanopore with the portable MinION device [66, 67] and the newly released PromethION. This technology was rather unpredictable in terms of output-quantity and error rate (which was higher than PacBio), but has recently been advancing up to an accuracy of 85% [68] and an output of 490Mb [69], with expectation of 90% and 2GB to come. Tools such as LINKS [70], NaS tool [71] and Nanocorr [69] already allow allow scientists to harvest the potential of this new technology in genome assembly.

Finally, the Moleculo platform offers synthetic long reads by fragmenting the genome into large 10kb segments, barcoding each segment, and sequencing them using regular Illumina technology. The short reads are then assembled into a small amount of synthetic, high-fidelity, longer reads. The application is especially geared towards haplotyping [72] or completing near-finished assemblies.

Currently, hybrid assembly is most often employed. With advances in SMRT technology (reduced error rate, longer read lengths) and assembly algorithms (lower memory requirements), this long read technology is capable of delivering near-complete assemblies, or start-to-end transcripts (in the case of transcriptome sequencing). Even highly repetitive or heterozygous genomes will be assembled, especially when combined with optical maps. The days of short reads will forever be over...

2.1.2 Masking the genome

Before locating genes in the genome sequence, it is best practice to filter out any repeat sequences. This will avoid predicting transposable element (TE) proteins, and help the prediction software by removing potential hurdles. Homology-based repeat finding methods such as RepeatMasker [73] rely on curated libraries of known repeat family consensus sequences (e.g. RepBase [74]). *De novo* or *ab initio* methods analyse the genome sequence itself to build custom repeat libraries. Popular *de novo* repeat finders are RepeatScout [75], ReCon [76] Red [77] and RepARK [78]. The latter does not require a consensus genome sequence but focuses on the NGS reads to build the repeat library. Finally, consensus methods combine results from multiple complementary methods (e.g. RepeatModeler [79] or REPET [80]), where after repeat element classifiers annotate the results (e.g. PASTEC [81], part of the REPET package, and TEclass [82]).

Transposable elements pose specific problems for gene prediction. Inadequate detection and filtering can lead to the disruption of non-TE genes (f.e. the presence of a TE within a gene intron) or 'over-prediction' (i.e. predicting too many genes by including repeat proteins). The presence of repeat proteins might also influence the training of *ab initio* models, which ultimately impacts the entire gene prediction. Relying purely on the detection of known families is not feasible for unknown species, because species-specific repeats and novel TE classes will remain absent from the databases. Additionally, the difference between different detection methods – and their sensitivity – can substantially increase or decrease the detected TE content. In *Zostera marina (chapter 6)*, the RepeatModeler approach resulted in a genome coverage of 45%, while a more sensitive REPET run covered nearly 63%. Such differences also illustrate the complexity of TE detection and annotation, something the community is trying to resolve with calls for proper benchmarking [83].

2.1.3 Structural Annotation: gene prediction

Prokaryote genes structures are usually very simple and defined by open reading frames (ORFs). There are few repeats and genomes are very gene-dense, which occasionally leads to overlapping ORFs. In eukaryotes however, genomes are less gene-dense and contain many more repeats. Additionally, coding regions (exons) are often interspersed with long non-coding intervening sequences (introns). This complex eukaryotic puzzle of exons and introns is even more difficult to solve when introducing alternative splicing, which allows a single gene sequence to produce multiple transcript variants. Many different gene prediction algorithms have been developed to tackle this issue and cope with the ever increasing amount of data. Lately, rather than developing new methods, more effort is being put into 1) maximum integration of NGS data (GeneMark-ET [84], CodingQuarry [85]), 2) combining/integrating different prediction results and producing consensus models (JIGSAW [86], Evigan [87], Evidence Modeler [88], iPred [89]), and 3) developing efficient annotation pipelines that automatically build training sets (GeneMark-ET [84], MAKER2 [90], SnowyOwl [91], BRAKER1 [92]). A recent overview of current methods for automated annotation can be found in Hoff & Stanke [93].

Content and Signal Sensors

While each prediction software has its own way of integrating the different data sets, all rely on two types of 'sensors' within the DNA sequence to accurately delineate gene structure and organisation. The first type, a 'content sensor', models variable-length features and classifies the DNA into different types (coding or non-coding i.e. intron, UTR, intergenic). Extrinsic content sensors make use of the similarity between the genome-of-interest and a protein or nucleotide sequence (Expressed Sequence Tags, SwissProt proteins, the genome itself, or another genome using the 'conserved exon method' [94]). Intrinsic content sensors try to define the innate characteristics of the sequence itself, by modelling hexamer frequencies (Mathe et al., 2001), nucleotide composition, codon usage, base occurrence periodicity and GC content. The models that scan for these properties are often Markov Models (MMs), or derivatives thereof *(Table 2.1)*. It must be noted that the intrinsic methods rely on extrinsic data for training.

A 'signal sensor' models features of fixed length. These features (splice sites, START site, STOP site, TATA box,...) most often indicate a change in the DNA-type (e.g. splice donor: changing from coding exon to non-coding intron; STOP site: change from coding exon to non-coding UTR or intergenic region). A signal sensor often

A **Markov chain Model** is the simplest form of the **Markov Model**. In the context of gene prediction, it assumes that the probability of a particular nucleotide (A, C, T or G) occurring at a given position is entirely dependent on the previous *k* nucleotides. The higher the order – *k* – of the MM, the more history the model contains and the higher its predictive value.

Unfortunately, a high-order MM requires a vast amount of training sequences to estimate the probabilities (a k^{th}-order MM requires 4^{k+1} probabilities). A second disadvantage of the high-order MM is the reliability of its probability estimates: larger k-mers occur less frequent. The most popular Markov Models are listed below.

Content sensors	
homogeneous MM	each position in the sequence is scored with the same model
inhomogeneous MM	different models for different positions in the sequence
Interpolated Markov Model (IMM)	The IMM linearly combines probabilities from contexts of varying length (i.e. k-mers that occur frequently are given high weights). The lower-order models (more data available) help to smooth the predictions of the higher-order models and improve performance. First pioneered in GLIMMER [95]
3-periodic MM	the most commonly used inhomogeneous MM in gene prediction, pioneered by GENMARK [96]. Each codon position has its own model, as well as an additional (homogeneous) model for the non-coding DNA. Each position of the genome is therefore evaluated by no less than 7 different models (3 coding models, both in forward and reverse strand, and 1 non-coding).
Interpolated Context Model (ICM)	the markov chains that make up the IMM do not have to exist out of all *k* adjacent bases. Instead, only the positions that are most informative, are chosen. The motivation behind this scheme states that certain positions within the *k* previous nucleotides are irrelevant (e.g. the third codon position often does not influence the amino acid translation) [97].
Signal sensors	
Position Weight Matrix (PWM)	An inhomogeneous zero-order MM that assumes each base occurs independently with a given frequency. A simple probability matrix describes the frequency of each base. These simple models are often used for non-coding regions and signals such as transcription-factor binding sites.
Weight Array Matrix (WAM)	an inhomogeneous higher-order PWM, represented as an array of Markov Chains. Often used for splice sites [98].
Windowed WAM (WWAM)	the probability at each position is averaged over neighbouring positions [99].

Table 2.1. Widely used Markov Model content and signal sensors in gene prediction. Adapted and modified from [100]

employs a PWM or WAM (e.g. SpliceWAM plugin for EuGene [101]) *(Table 2.1)*, or even a Support Vector Machine (SVM) e.g. SpliceMachine [102]. The latter is used for splice site prediction in all genome projects presented in this thesis. SpliceMachine is a supervised machine learning classifier that defines a set of features that best represent bona fide splice sites. These features include positional information (nucleotide preferences), compositional information (oligomer preference) and codon potential (reading-frame-based codon bias). This high-dimensional local context allows the SVM to classify potential splice sites and assign a score ranging from 1 (true splice site) to -1 (false splice site).

A Hidden Markov Model (HMM) [103] is another popular signal model that is able to accommodate some variation in the signal length as it allows insertions and deletions. This model will try to assign all nucleotides into specific states (e.g. exon, donor, intron, acceptor). Each state has its own emission probabilities (models base composition with a Markov Model), as well as transition probabilities (i.e. the probability of moving from

one state to another). When the HMM scans the sequence, each nucleotide is assigned to a state, generating a hidden 'state path'. This state path is a Markov Chain, as the state we are in depends entirely on the previous state (an intron state nucleotide can only be preceded by a donor state nucleotide). The HMM will find multiple possible paths, each with its own score owing to the emission and transition probabilities. The HMM will try to find the most optimum path thanks to dynamic programming algorithms such as Viterbi.

Combining evidence

While the HMM example described above illustrates a rather small implementation (4 states for splice site detection), it can easily be scaled up to generate a gene-finding HMM with many states. A General Hidden Markov Model (GHMM) provides a framework that separates the overall HMM structures from the embedded submodels, meaning that new models can easily be added, or existing model retrained [104]. Additionally, the GHMM can incorporate constraints on different levels such as frame constraints – the coding sequence (CDS) needs to maintain a correct reading frame – and length constraints (length distributions of exons/introns/... are taken into account). To cope with the amount of possible paths (~gene structure), most gene finders represent the signal and content sensor evidence on a graph *(Figure 2.2)* and traverse this using dynamic programming algorithms [101, 105].

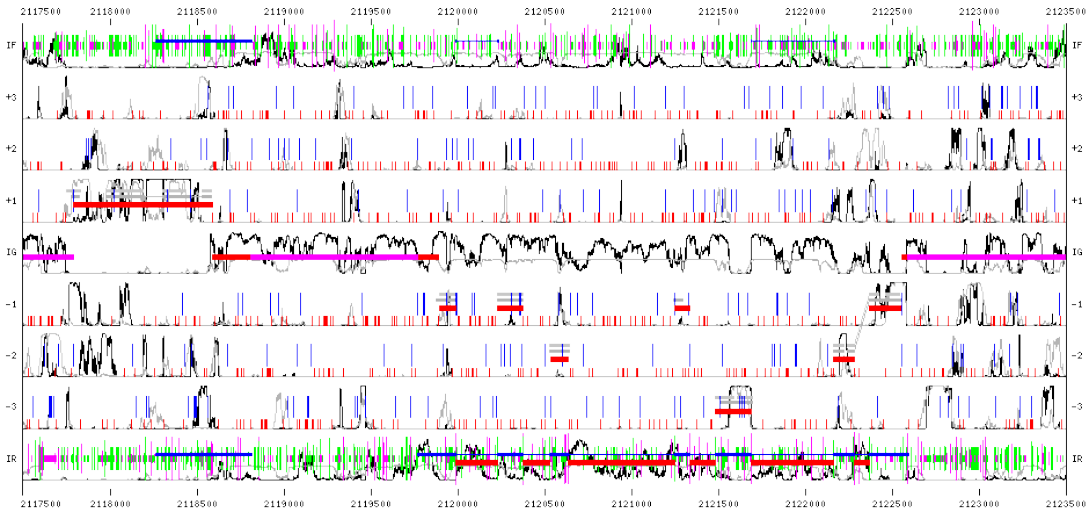


Figure 2.2. The EuGene undirected acyclic graph. The graph displays 9 tracks. The top and bottom tracks display donor (green) and acceptor (pink) scores for the intron-forward (IF) and intron-reverse (IR) strand. The middle track (IG) models intergenic regions. The 6 remaining tracks model possible start (vertical blue) and stop (vertical red) codons as well as the nucleotide IMM score (black) and protein MM score (grey). Two predicted gene models are displayed, showing the exon structure (horizontal red bars in the [+3+2+1-1-2-3] tracks) and the introns (horizontal red bars in the IF and IR tracks). Extrinsic evidence is also visible: EST alignments (blue bars in IF and IR track) and protein alignments (grey bars in [+3+2+1-1-2-3] tracks). The genomic region displayed here is *Zostera marina* scaffold_1 (positions 2117500–2123500).

Gene prediction pipelines are geared towards integrating and combining different evidence sources before traversing the graph. AUGUSTUS accepts 16 different types of 'hints' (start, stop, exon, CDS,...) which are fed into the program via a General Feature Format (GFF) file [106, 107]. Parameters are manually set for each evidence source, or optimized through iterative training. Alternatively, EuGene provides a modular system with different plug-ins that allow integration of any data type [101]. The parameters for each module are determined using an optimization scheme involving a genetic algorithm. Different individuals (~parameter combinations) are evaluated and assigned a fitness score i.e. how well does the gene prediction, resulting from this parameter combination, perform in comparison to the 'golden standard' training set. Afterwards, the fittest individuals from that generation give birth to a new generation, with some modifications (crossovers, mutations). After a few generations, the algorithm usually finds the best parameter combination. This unique algorithm and versatility in data usage through the plug-ins, provides great flexibility in the gene prediction

pipeline. Downsides are the steep learning curve and obvious difficulty in managing the pipeline whereas other software (e.g. AUGUSTUS, GENEMARK, MAKER) are more user-friendly, but generally lack flexibility in data integration.

The annotation pipeline

An example structural annotation process is outlined in *Figure 2.3*. First, all available extrinsic data is mapped onto the genome-of-interest. The alignments of ESTs, RNAseq, proteins and nucleotide sequences provide information on putative splice sites and (non-)coding regions. The intrinsic content and signal sensor models of the prediction software could be trained from this information alone, providing many data points, some of which could be false positives (due to the nature of the mapping algorithms and the thresholds involved). An alternative route is to feed the information to an initial prediction software (e.g. GeneMark-ET [84] or AUGUSTUS [106, 107]) to generate a first annotation draft, possibly succeeded by a manual curation in a genome browser (e.g. GenomeView [108]). In the end, we obtain a training set that contains less data points, but is of higher quality. The training set can be used to 1) train the intrinsic content and signal sensor models (if not already done so as mentioned above), and/or 2) optimise the parameters of these models. If required (and enough data is available) the training set can be split into two sets, one for training and one for evaluation (how well does the trained prediction result fit to the evaluation set?). Once all models are trained and optimized, the genome-of-interest is ready to be annotated. Right before we run the gene prediction, we again provide the extrinsic data together with repeat information to the software (EuGene [101]). The gene prediction is often rerun several times due to the fine-tuning of certain parameters or the inclusion/exclusion of certain data (e.g. TBLASTX data can be really messy and is sometimes discarded). If needed, the annotated genome can go through a post-processing phase, involving integration of other gene types (RNA genes, pseudo genes) and/or manual/automatic error correction using a web interface like WebApollo [109] or Online Resource for Community Annotation of Eukaryotes (ORCAE) [110]. In ORCAE, expert annotators are able to look at individual genes together with all data sources and perform structural (and functional) corrections, a time-consuming process that nevertheless results in high-quality annotations.

RNA genes

It must be pointed out that the previous sections, describing the annotation process, all focus on protein-coding genes, the main focus of genome projects in line with the central dogma of gene expression. Additionally, highly-abundant and functionally important non-coding RNA genes such as transfer-RNAs (tRNAs), ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs) will also be identified. Predicting such RNA genes is mostly a homology-based procedure. Programs like Infernal [111] make use of the co-variance models (CM; a multiple sequence alignment complemented with structural features) from the RFAM database [112]. tRNAscan-SE [113] also employs a CM to model tRNAs, while RNAmmer [114] relies on a library of ribosomal RNA HMMs.

Long non-coding RNAs (lncRNAs) are a large and diverse class of transcribed RNA molecules that have little to no protein-coding capability and extend beyond a length of 200 nucleotides. They play a vital role in the regulation of many cellular and developmental processes, and are crucial regulators of gene expression [115]. These lncRNAs have fewer exons on average, but exhibit the same canonical splice site signals and alternative splicing tendencies as mRNAs [116]. But how do we predict such ncRNA genes? Identification of chromatin signatures, more specifically K36-K4 chromatin domains, and the correlation with transcriptomics data makes it possible to discriminate between biologically significant lncRNAs and transcriptional noise [117]. The sequence itself also provides clues: the absence of coding potential, and presence of RNA secondary/tertiary structures proves valuable for identifying lncRNAs [118].

Micro-RNAs (miRNAs) constitute a class of smaller (~22 nucleotides) non-coding RNAs that play an important role in the regulation of gene expression at the post-transcriptional level. The primary miRNA transcripts are processed by RNAse III enzymes (Drosha and DICER) to form a mature 22nt miRNA. They are subsequently integrated into a ribonuclear particle to form the RNA-induced Silencing Complex (RISC), which mediates gene silencing. Transcriptional noise combined with their small size makes computational identification very difficult, often requiring experimental validation [119].

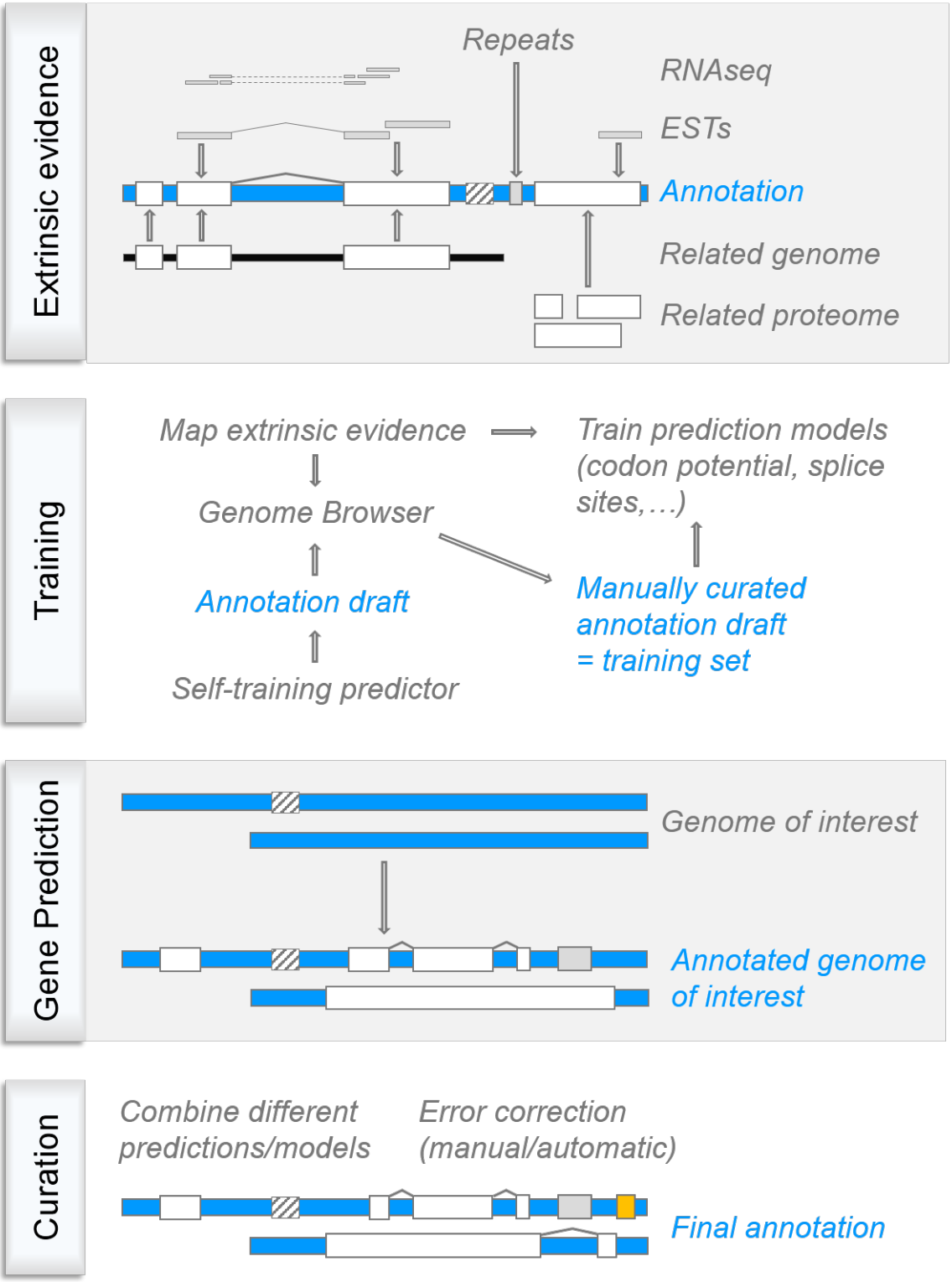


Figure 2.3. A typical annotation process. (Extrinsic Evidence) The comparative resources available when starting the annotation process. **(Training)** The combination of extrinsic evidence and quick-and-dirty predictions allow an expert annotator to construct a golden standard (i.e. the training set) that is then used for training the intrinsic models of the prediction software. **(Prediction)** After training, we map all available extrinsic data on the entire genome-of-interest and unleash the prediction software in order to obtain an annotation. **(Curation)** Gene models from other prediction programs are often added (e.g. RNA genes) and afterwards all models can go through an automated/manual cleaning process, resulting in the final annotation. Colour codes: blue (genome sequence), white (exon), grey (repeat), yellow (RNA gene) and white-black latching (sequence gap). Arches indicate introns.

Pseudogenes

Pseudogenes are difficult to predict from an algorithmic point of view because certain biological assumptions are hard-coded into the prediction software e.g. the coding sequence does not contain internal stop codons, or frameshifts. In such cases, prediction software will try to circumvent the disruptive mutations by introducing additional introns that allow the coding sequence to maintain the correct reading frame, or simply by breaking up the gene into two separate structures. Often, alternative methods are required to properly identify and annotate pseudogenes (PseudoPipe [120] and PseudoDomain [121]).

2.1.4 Functional Annotation

The main objective of a genome project is to get to the core of a specific organism by understanding what the genes do and how they work together. One of the best ways to achieve this is by identifying the exact function of each gene. Because protein-coding genes make up the majority of the gene repertoire, they are the aim for most analysis tools. Experimental methods involve wet lab approaches to functionally analyse a single gene, generating a very low throughput. Computational approaches use a combination of amino acid properties, sequence structure, homology, motifs and domains to predict the sub-cellular location, enzyme function, signal peptide, transmembrane topology and involvement in pathways [122–124]. Popular functional analysis tools in genome projects are Phobius [125], InterProScan [126], BLAST2GO [127], EggNog [128, `http://eggnogdb.embl.de`], SIFTER [129] and PANNZER [130]. Even with a wealth of data and many functional annotation tools to our disposal, the majority of predicted genes will still end up with the label 'hypothetical protein' or 'uncharacterised protein' [131] and will require additional attention [132].

Within the BEG group, we try to go beyond the assignment of GO-categories and protein domains, and aim to provide clear one-line human-readable functional descriptions for all genes. The Online Resource for Community Annotation of Eukaryotes (ORCAE) platform displays every source of information for a given gene, allowing expert annotators to assess and decide on the function [110]. Additionally, functional descriptions will also be transferred from trusted sources (e.g. SWISSPROT) via homology, or automatically inferred from different text sources (e.g. in-house text-mining tool AnnoMine). Providing a list of domains and categories might be interesting from a computational point of view, but does not provide great value for wet-lab biologists.

2.1.5 Challenges in Genome Projects

Genome projects face many challenges. Genomes can be extremely large (e.g. *Paris japonica*: 150Gbp [133]) and complex, with high levels of ploidy, heterozygosity and repeats. Additionally, the presence of phenolics and polysaccharides within (plant) tissues complicates the extraction of large quantities of high-quality DNA, a requirement for today's sequencing technologies. Furthermore, obtaining pure DNA samples is difficult. Often, DNA from (genetically) different individuals needs to be pooled in order to obtain the quantities needed for sequencing. The presence of bacterial contaminations and over-representation of organellar DNA also dilutes the sample DNA. On top of such problems, genome projects today strive towards sequencing 'sexy' exotic – and complex – species with minimal amounts of funding (e.g. the $1,000 genome) and time. This sexiness might be coupled with previously undocumented – and thus harder to predict – features. Such issues have a serious impact on the assembly and annotation steps within the genome project. They also make the downstream analysis more difficult and make it harder for scientists to answer the question that kick-started the genome project: what are the (unique) properties of this genome? Next up I will discuss several ways to cope with this reality.

Heterozygosity

A diploid organisms contains 2 copies of each gene. These copies can be identical (homozygous) or variable (heterozygous). If a heterozygous diploid organism is sequenced, the allele variants introduce additional complexity in the OLC and DBG assembly graphs. The heterozygous regions produce imperfect overlaps and different k-mers, which can lead to strange assembly artefacts. Different strategies have been developed to

counter these problems. The first approach is to just discard one of the similar sequences from the graph bubble (e.g. Velvet). A second more elaborate and time-consuming option is the development of inbred lines to reduce the level of heterozygosity to a bare minimum. A last option is to develop assembly software able to cope with this data such as Hapsembler [134], dipSPAdes [135] and Platanus [136], as well as the commercial platform DeNovoMAGIC^{TM} (`http://www.denovomagic.com`), a haplotype-aware assembler geared towards complex crop genomes. Depending on the level of diversity between alleles, other options are long reads and assembly-refinement pipelines (HaploMerger [137], ScaffoldScaffolder [138] and Redundans [139]).

Assembly contamination

Removing bacterial contaminants from eukaryotic assemblies is an absolute must. They influence statistics and might provide a distorted view on the actual gene content of an organism (e.g. influence on gene family analysis). Several of the genome projects mentioned in this thesis had to cope with bacterial contamination in the assembly process. The contamination has a detrimental effect on the sequencing outcome, with an over-representation of foreign contaminant DNA that needs to be removed either before assembly (read level) or after assembly.

Contaminant screening on the read level mostly involves filtering out sequencing adapters and well-known contaminants, leaving most bacterial DNA in the sample. Only a handful of tools exist to filter out contaminants in NGS reads such as QC-chain [140]. Screening of (draft) assemblies often involves GC content filtering and taxonomic identification [141, 142], but both methods have their pitfalls and should be used with caution. GC content filtering allows for a very fast approach but could result in the dismissal of 'true' host contigs that display local variation (e.g. GC islands, outlier chromosomes, organellar DNA). Similarly, contigs originating from bacteria with roughly the same GC content as the host would be retained. In *Zostera marina*, the assembly contained bacterial and archaeal sequences, even after extended GC filtering. Taxonomic identification relies on a best-hit approach in relation to public annotated databases, assigning a taxonomic id to each contig. While this approach is slower, it is more accurate and allows for a more fine-tuned comparison. The downside is its reliance on homology. If genes have no homologs (e.g. species-specific genes), or if there is no comparable content (e.g. contigs existing entirely out of repeat elements), then taxonomic identification of said contig is impossible. Additionally, if multiple genes have a bacterial origin due to horizontal gene transfer (e.g. Small Outlier Chromosome), assigning the correct (non-bacterial) tag to the contig will be difficult. While these approaches yield excellent results, there are no standardized protocols that automatically sort everything out at 100% accuracy. Contaminant filtering will require some manual intervention and a case-by-case configuration.

The approaches described above are of absolute necessity for marine eukaryote genome projects, which often have difficulty establishing axenic cultures (i.e. only a single species or strain) and exhibit bacterial and fungal contamination. The sporadic use of seawater in absence of a well-defined growing medium adds to this complexity. However, it is highly likely that the interaction with several of the contaminants is beneficial and required for its culturing [143]. Hence, complete contaminant removal will remain impossible.

Quality Measurements

Before being published, genome projects are required to provide statistics to illustrate the quality and completeness of an assembly and annotation. Assessing the quality of an assembly is usually done through sequence-level metrics such as l50 and n50. When sorting all assembly contigs by size (from largest to smallest), l50 reports the number of contigs required to contain 50% of the genome, while n50 reports the size of the smallest contig within that l50 set. Additionally, the total assembly size is compared against size estimates (usually by Pulsed-Field Gel Electrophoresis). In order to compare different assemblies, detailed visual reports are generated, including cumulative length distributions and GC% plots (QUAST/metaQUAST [144, 145], Hawkeye/AMOS [146] and REAPR [147]). In addition to sequence-level metrics, a genic level approach can substantially enhance assembly quality assessments. Tools such as CEGMA [148] and BUSCO [149] have defined a set of conserved genes present in a wide range of eukaryotes and try to locate these genes in (draft) assemblies, providing a measurement for 'genic space' capture.

Evaluation of popular genome assembly software has revealed many common errors such as contig misjoins, duplicated contigs and compression of repeats [150]. While quality assessment tools provide detailed and visual reports, genic-level approaches can indicate the completeness of our (draft) assembly. Clearly, it should be standard practice to include such measurements in today's genome papers.

The quality of an annotation is traditionally reported in the number of genes supported by extrinsic evidence: ESTs, protein homologs and RNAseq. Higher evidence percentages indicate a compliance with said evidence but do not necessarily prove that the annotation is correct e.g. even if all introns of a gene are supported by ESTs or RNAseq, there could be other exons (and introns) that aren't annotated. Another measurement is the total amount of predicted genes and its comparison with already-annotated related species. Too low a number indicates possible 'under-prediction' (i.e. the gene prediction pipeline did not pick up all genes), too high a number could indicate either 'over-prediction' (i.e. predicting genes where there aren't any to be found) or possible genome duplication(s). Similarly, gene family analysis can also highlight these issues (e.g. too many single-copy genes could point to over-prediction), as do genic-level approaches (see previous paragraph). Occasionally, we visualise the ratios between the predicted proteins and their respective top BLAST hit. The rationale behind this method is that protein length is conserved throughout evolution, and homologs should have similar lengths, resulting in a normal distribution.

2.2 Marine Life

The ocean covers 71% of the Earth's surface, yet an estimated 95% remains unexplored. Organisms inhabiting this vast world are vital for key ocean services such as the regulation of Earth's climate by absorbing excess heat and CO_2 from the atmosphere, thereby buffering the greenhouse effect, recycling nutrients (phosphor, nitrogen, sulfur and iron) and the production of the bulk of the Earth's oxygen (ranging from 50% to 70% depending on the source). However, the dynamics of marine ecosystem are being altered by climate change (warming, ocean acidification, deoxygenation, and altered food inputs), pollution and the resulting population effects (e.g. harmful algal blooms). This inevitably impacts not only marine ecosystems, but the entire planet, prompting us to take adequate measures to ensure the oceanic ecosystems are protected [151]

To improve our knowledge on global marine life, several ocean sampling expeditions have been launched in the last decade such as the Sargasso Sea study [152] and the Global Ocean Sampling [153] and TARA Oceans [154] initiatives. They focus primarily on plankton and report an extraordinary diversity with a vast majority of unidentified species, a multitude of interactions, and the environmental factors that influence population dynamics. Understanding how these marine organisms evolve and adapt is key into understanding the ecosystems that fuel our planet. Why are specific organisms so successful in certain ecological niches? Which adaptations help them thrive? Why do certain organisms interact within an ecosystem? The answer to such questions lies in the biology of the organisms, and more specifically in the genome that encodes the organisms properties.

Ideally, a reference database of marine genomes with adequate representation of all taxons would be a great starting point for scientific research. While more and more marine genomes are being sequenced, we are still far from this ideal picture. Within the next sections I will introduce the marine organisms that were chosen to be sequenced, annotated and analysed, and contribute to this growing repository of marine data.

2.2.1 Prasinophytes & the Mamiellales

The green lineage (Viridiplantae) contains two main divisions. The Chlorophyta contain most extant green algae, while the Streptophyta group freshwater green algae ('charophytes') and the land plants [155]. One of the earliest diverging branches within the Chlorophyta are the prasinophytes [156], a group of mainly marine planktonic species that gave rise to the core chlorophytes (Pedinophyceae, Chlorodendrophyceae, Trebouxiophyceae, Ulvophyceae and Chlorophyceae) *(Figure 2.4)*. To understand the nature of the last common ancestor of all green plants, studying basal green algal lineages (in both Chlorophyta and Streptophyta) can provide many insights [157]. According to a study on *Mesostigma viride*, an early offshoot of the Streptophyta branch (and a former member of the prasinophytes) [158], the most ancestral green flagellate is an asymmetric cell with two flagellae, an eyespot (photoreceptive organelle) and a layer of scales [159].

The prasinophytes are a diverge assembly of unicellular organisms with a wide range of cell shapes and sizes, flagellae and scales. Some have no scales, or no flagella, or neither (e.g. *Ostreococcus*). Phylogenetic analysis of 18S rRNA resulted in the identification of seven independent monophyletic prasinophyte lineages or clades [160]. Afterwards, two additional clades were added that lacked any cultured representative and consisted entirely out of environmental sequences [161]. All nine lineages differ vastly in morphology, life cycle and ecology [162, 163] *(Figure 2.4)*. When comparing the phylogenetic results to the proposed taxonomic hierarchy, some clades correspond to existing orders or classes (clade II: Mamiellophyceae, clade III: Nephroselmidophyceae, clade IV: Chlorodendrophyceae) [164]. The taxonomic classification of prasinophytes – and green algae in general – has experienced many changes over the years, and will continue to do so until all clades can finally be resolved.

Within the prasinophytes, the class of the Mamiellophyceae (clade II) [166] contains three main orders of aquatic eukaryote green algae: the Dolichomastigales, the Monomastigales and the Mamiellales. The latter contains two families, the Bathycoccaceae (*Ostreococcus* and *Bathycoccus*) and the Mamiellaceae (*Micromonas*, *Mantoniella* and *Mamiella*). The definition of Mamiellophyceae (Marin et Melkonian classis nova) is listed below [166]. The Mamiellophyceae are said to be highly important in coastal and polar ecosystems and are amongst the most ecologically successful picoeukaryotes in the ocean [166].

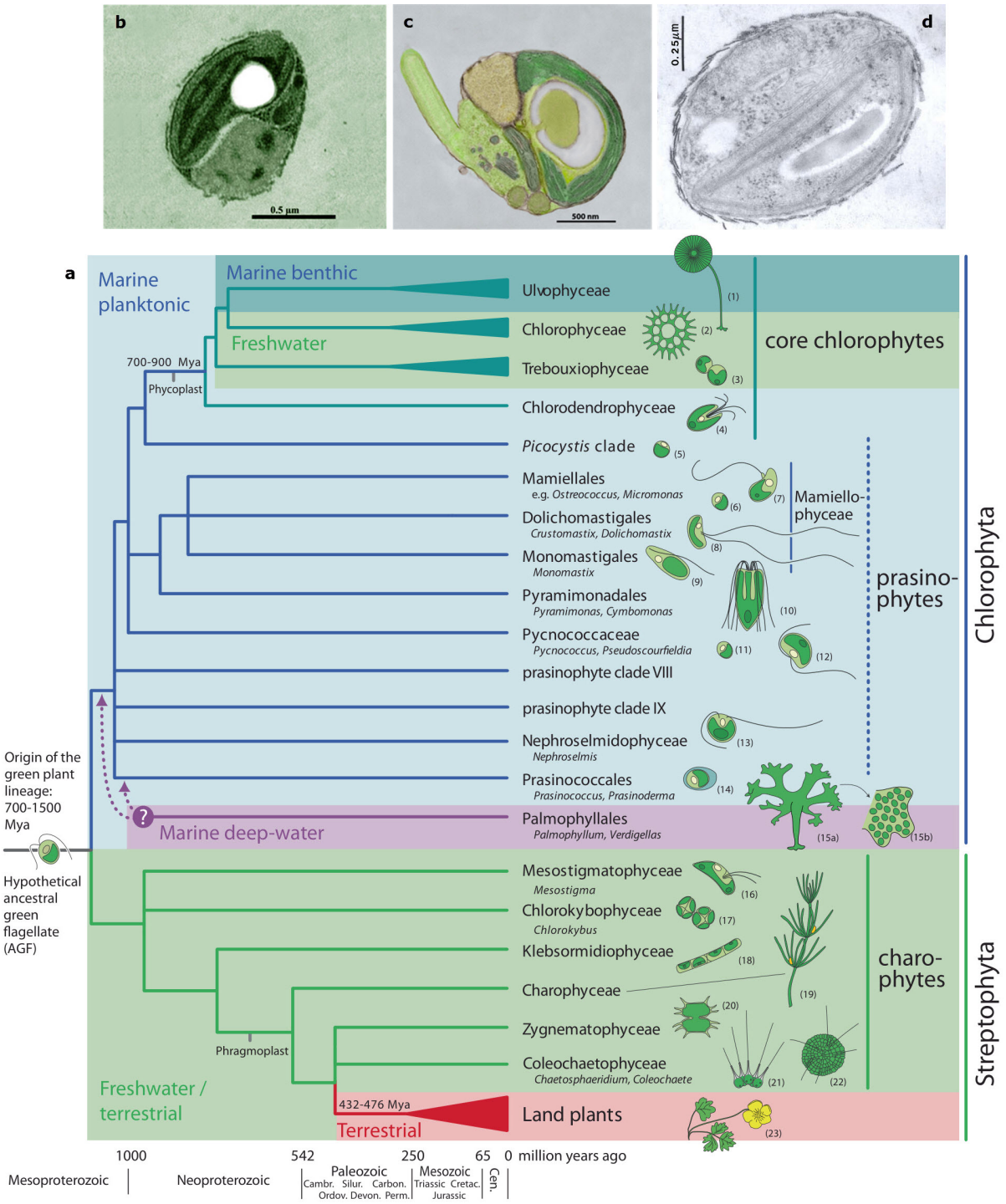


Figure 2.4. Phylogenetic relationships among the main lineages of green plants. (a) The tree topology is a composite of accepted relationships based on molecular phylogenetic evidence. Uncertain phylogenetic relationships are indicated by polytomies. The divergence times are rough approximations based on the fossil record and molecular clock estimates. These age estimates should be interpreted with care as different molecular clock studies have shown variation in divergence times between major green plant lineages. Drawings illustrate representatives of each lineage: (1) *Acetabularia*, (2) *Pediastrum*, (3) *Chlorella*, (4) *Tetraselmis*, (5) *Picocystis*, (6) *Ostreococcus*, (7) *Micromonas*, (8) *Crustomastix*, (9) *Monomastix*, (10) *Pyramimonas*, (11) *Pycnococcus*, (12) *Pseudoscourfieldia*, (13) *Nephroselmis*, (14) *Prasinococcus*, (15) *Verdigellas* (a: general habit, b: individual cells in a gelatinous matrix), (16) *Mesostigma*, (17) *Chlorokybus*, (18) *Klebsormidium*, (19) *Chara*, (20) *Xanthidium*, (21) *Chaetosphaeridium*, (22) *Coleochaete*, (23) *Ranunculus*. Source: Leliaert *et al.* [162]. **(b)** *Ostreococcus tauri*, photo courtesy of Hervé Moreau (Laboratoire Arago). **(c)** *Micromonas pusilla*, photo sourced from `http://www.mbari.org/`. **(d)** *Bathycoccys prasinos* sp. nov. Source: [165]

> ”Eukaryotic algae, growing in water. Cells usually solitary, with 2 flagella (equal to subequal, or unequal), or a single flagellum, or lacking a flagellum. A single chloroplast, surrounded by two membranes, with chlorophylls a and b, nearly always with prasinoxanthin. Cells sometimes with two chloroplasts. Eyespot posterior, or lacking. Cells and/or flagella covered by scales in 1-2 layers, or without scales. Scales flattened, rounded to elliptical, mostly ornamented like a spider web with concentric ribs, or uniformly reticulate. Cells and flagella lacking an inner layer of small square scales. Algae predominantly inhabiting marine water, but also freshwater. The first pair of Helix E23_12 in the nuclear 18S rRNA is G-U instead of A-U.”

Many Mamiellales species are cryptic species, meaning that two or more species are hidden under the same species name. This is often the case when the species are closely related and cannot be distinguished from each other based on their morphology (one ’morphospecies’) or, in a more simple terminology, they look alike but are genetically quite diverse. As a result, genetic analysis has revealed several independent lineages within Mamiellales species [167–169]. In the next sections, a more detailed description will be provided of several Mamiellales species.

2.2.2 *Ostreococcus*

In 1994, a photosynthetic picoeukaryote is discovered in the marine Mediterranean Thau lagoon, France [170]. Cell measurements revealed that this green alga, named *Ostreococcus tauri*, is the smallest eukaryote yet described. The organelles and ultrastructure of the cell advocate placing *O. tauri* in the Prasinophyceae [171]. The naked non-flagellated cells have a very elementary organisation, with a relatively large nucleus, single mitochondrium, starch granule, Golgi body and a very reduced chloroplast [171]. Further phylogenetic analysis confirms its placement in the Mamiellales order, while Pulsed-Field Gel Electrophoresis (PFGE) determines the nuclear genome size [172]. With 14 chromosomes totalling an estimated size of ~10.2 Mb, *O. tauri* is both the smallest eukaryote in cell size and genome size.

Why such small cell and genome size? Which forces have lead to this densely packed genome [173]? According to theories by Cavalier-Smith [174] and Patrushev & Minkevich [175], an organism (e.g. *Ostreococcus*) would reduce its genome size to get rid of the DNA disadvantages, and reduce its cell volume to maintain a low doubling time and high growth rate, a low morphological complexity and a high surface-to-volume ratio. All these changes are made to optimize the occupation of a certain ecological niche as efficient as possible (~nutrient availability and uptake). Modelling approaches do indeed confirm that extreme small cell size in phytoplanktonic organisms leads to a trade-off between cell size, nutrient and light affinity, and growth rate, which arises due to the necessary allocation of resources to nonscalable structural components (e.g. cell membrane minimal thickness) [176, 177]. *Ostreococcus* strains can be divided into four clades – ecotypes – based on a maximum likelihood phylogeny inferred from rDNA regions [169, 178] *(Figure 2.5)*, and related to their adaptation to light intensity [179] and geographical location [180]. Clade A contains surface strains that are adapted to high light intensities (e.g. *Ostreococcus lucimarinus* [181]). Clade B is comprised of strains living in low light intensities at the bottom of the euphotic zone (e.g. *Ostreococcus* sp. RCC809, a clone of RCC141). Clade C contains light-polyvalent strains such as the lagoon reference strain OTH95 [182], while clade D contains low-light strains (e.g. *Ostreococcus mediterraneus* [169]).

The genome sequence & structure

The creation of a Bacterial Artificial Chromosome (BAC) library allowed the sequencing of ~12% of the *tauri* genome which contained an estimated 1000 open reading frames [183]. A few years later, the entire genome was sequenced [182], totalling 20 chromosomes, a size of 12.56Mb and 8,116 predicted protein-coding genes, making *O. tauri* the most gene-dense free-living eukaryote known to date. The genomic structure revealed a peculiar heterogeneity, with two chromosomes (2 and 19) displaying a different organisation and function compared to the other chromosomes. Both aberrant chromosomes have a lower GC% and contain most of the 417 identified transposable elements (TEs). Additionally, chromosome 2 features a different gene codon usage and many small AT-rich introns. As a result, gene prediction is more complicated in this region, presumably leading to several prediction errors.

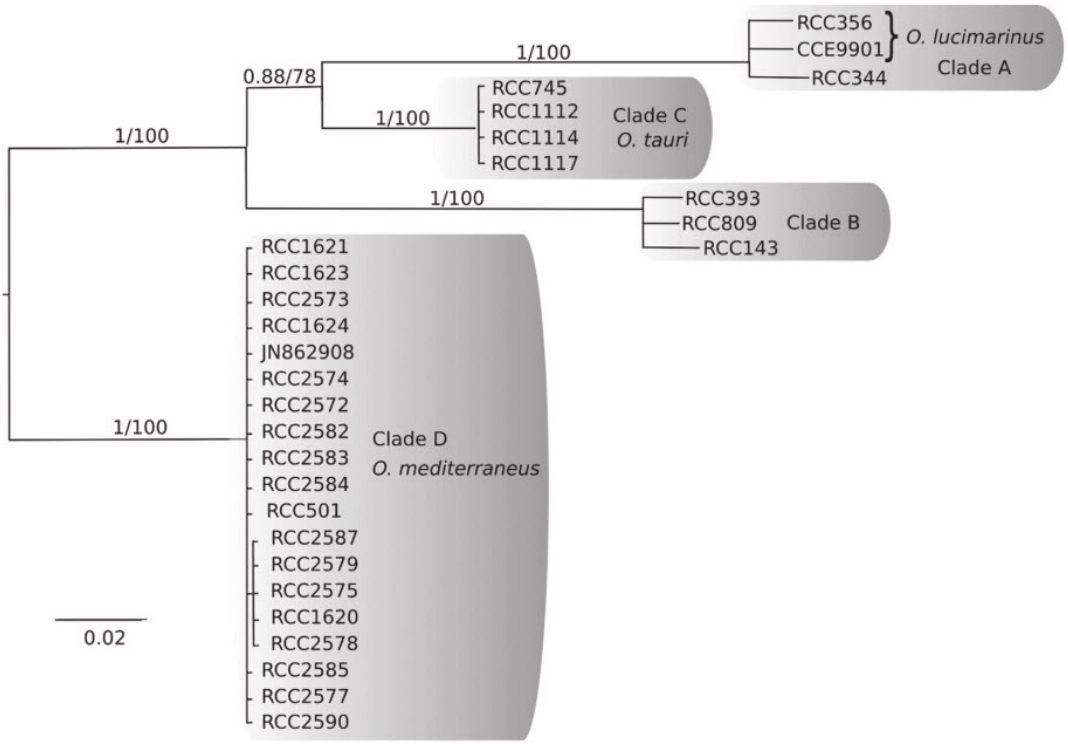


Figure 2.5. Phylogenetic relationships between *Ostreococcus* spp. clades in the Order Mamiellales. Phylogeny based on 18S rRNA genes. The branch numbers indicate posterior probabilities/bootstrap support values (in %). Source: [169]

While most eukaryotes have large intergenics region interspersed with many long repeat elements, 81.6% of the *O. tauri* genome sequence is taken up by coding sequences. The intergenics are shortened (average 196bp), most gene families are reduced, and certain genes are fused together. It must be noted that the genome sequence was never completely resolved. The many gaps that still remained lead to a partial state for several hundreds of genes.

Viruses

Ostreococcus cells are often infected by very large prasinoviruses (Phycodnaviridae) such as OtV5 (186,234 bp [184]), OtV-1 (191,761 bp [185]) and OtV-2 (184,409 bp [186]). The viruses are highly species-specific [187] and encode for a DNA polymerase, but not a DNA-dependent RNA polymerase. The prasinoviruses thus need to reach the nucleus in order to use the host RNA polymerase for gene transcription [188]. The viruses encode specific enzymes involved in several amino acid biosynthesis pathways (e.g. 3-dehydro-quinate synthase, involved in the synthesis of aromatic amino acids: phenylalanine, tyrosine, and tryptophan), a feature which is quite remarkable. The virus genomes evolve according to the 'genomic accordion' model: there is a balance between gene gains – mainly duplications, but also horizontal gene transfers from their hosts or other eukaryotic/prokaryotic sources – and gene losses, resulting in a relatively stable genome size [189].

O. tauri as a modelorganism

Ostreococcus tauri makes a good model organism because its is easy to culture (can be grown on just seawater) and has a fast growth rate and low generation time (they can divide 3–4 times each day). It is in a haploid phase, has a low cellular complexity and a small genome with a low amount of duplicated genes. The previous properties make the genetic manipulation of *O. tauri* much easier [190]. The next paragraphs will discuss the usage of *tauri* as a model organism in a variety of research fields.

Creating models of the ultrastructural complexity of entire eukaryotic cells is difficult. Nevertheless, combining electron cryomotography and *O. tauri* cells made it possible to produce high-resolution 3D models of an entire eukaryotic cell in near-native state [191] *(Figure 2.6a)*. The images showed remarkable results regarding cellular division, hinting that *O. tauri* might contain simplified mitotic mechanisms: 1) the nuclear envelope remained open during most of the cell cycle, 2) no condensed chromosomes nor a mitotic spindle were observed, and 3) the maximum number of observed microtubules was two, which is not enough to separate 20 linear chromosomes in a canonical fashion. A BUB1-like protein was identified, a protein kinase important for establishing the mitotic spindle checkpoint, meaning that some form of mitotic spindle is likely to exist [192, 193]. Further analysis revealed that mitotic *tauri* cells have an intranuclear heterochromatin-free 'spindle tunnel' with approximately four short and one long, incomplete microtubule at either end *(Figure 2.6b)*. This implies that chromosomes are most likely physically linked before initiating co-segregation [194], allowing a single microtubule to move more than one chromosome at a time (and mitosis to be completed in a single round of anaphase). Further analysis revealed that *O. tauri* chromatin is a disordered assemblage of nucleosomes, a 'polymer melt'. No large-scale reorganisations occur during mitosis because nucleosomes were rarely seen within the spindle tunnel. The centromeric nucleosomes could cluster in a ring surrounding spindle microtubules [195] *(Figure 2.6c)*.

While mitosis has been observed, meiosis (sexual reproduction) has not, although *tauri* possesses all necessary core meiosis genes [182], a pattern common to other Mamiellales species [196]. It is known that marine algae commonly suppress their meiosis capability while in culture, but indirect evidence of crossing-over and chromosomal segregation reveals that *tauri* populations can reproduce sexually [197].

Cyclin-dependent Kinases (CDKs) form heterodimers with a cyclin subunit to create CDK-cylin complexes that are able to regulate the cell cycle. Many plant genomes contain several copies of CDK and cyclin genes [198], while *O. tauri* shows a minimum, but complete, set of core cell cycle genes [199]. Thanks to its ability of natural synchronisation, the CDK expression can be studied throughout the cell cycle [200] which has led to the conclusion that the transcription of these cell cycle genes is under circadian control [201]. As with cell cycle genes, clock genes are also restricted to a minimal set with a conserved TOC1 (Timing of Cab expression 1) and CCA1 (Circadian Clock-Associated 1) gene [202], creating a simple two-gene oscillator clock that is very robust to light fluctuations [203, 204]. Even without transcription, the circadian rhythm persists, revealing a conserved non-transcriptional alternative mechanism involving oxidation–reduction circadian cycles of peroxiredoxins [205, 206].

A low-complexity green microalga is the ideal platform for the study of photosynthesis. Light-harvesting complexs (LHCs) are assemblies of different LHC proteins that bind chlorophylls and carotenoids. They collect and transfer solar energy to the photosystem reaction center. *O. tauri* possesses 1) an unusual prasinophyte-specific LHC protein type (in multiple copies), 2) the major LHCI proteins (single-copy), and 3) only two minor LHCII polypeptides (single-copy), supporting the theory that LHCI proteins arose first [207, 208]. Other unusual properties include a high number of DNA-repair enzymes that target UV damage [209] and the absence of protochlorophyllide reductase genes, meaning that chlorophyll can only be synthesized during the day much like angiosperms [182]. While the make-up of LHCs is extensively studied (involved in the 'light reactions' of photosynthesis), the subsequent 'dark light' reactions could also reveal interesting results as *Ostreococcus* has all the machinery necessary to perform C4 photosynthesis, a variant of classic C3 photosynthesis that is more efficient in CO_2-limiting conditions (such as algal blooms) [182].

Ostreococcus tauri as a bioremediator

Bioremediation is a technique often used in waste management. It employs organisms to remove or neutralize harmful pollutants from contaminated environments. Recent studies explore the use of *O. tauri* as both a biosensor and bioremediator. Sanchez-Ferandin *et al.* [210] describes the use of a luciferase-CDKA-fusion construct to test for toxic compounds, while Zhang *et al.* [211] explores the use of *tauri* in the biomethylation and volatilization of arsenic. Microalgae are well suited for these tasks and it is expected that more uses will be found for *tauri* in the field of ecotoxicology.

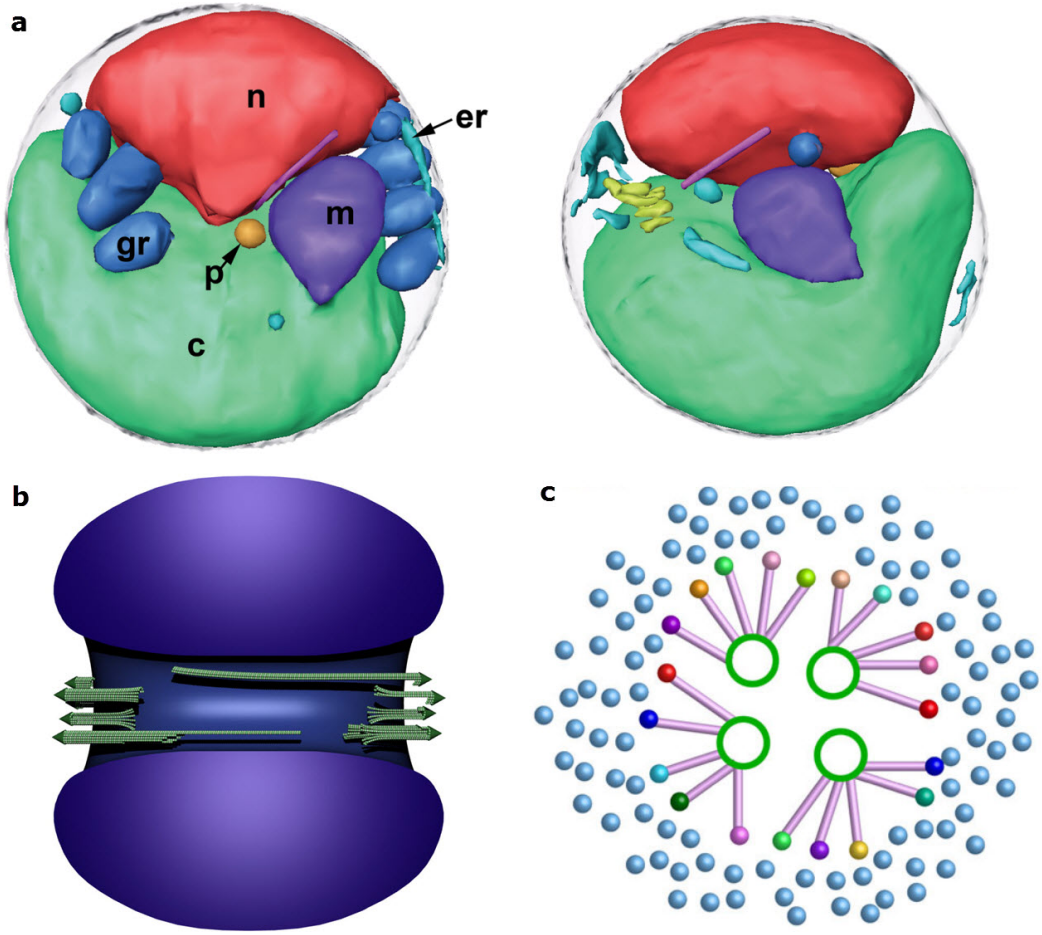


Figure 2.6. 3-D segmentations of two *Ostreococcus tauri* cells and a scale model of the spindle and chromatin configuration. (a) Shown here are the nucleus (n, red), chloroplast (c, green), mitochondria (m, dark purple), a Golgi body (yellow), peroxisomes (p,orange), granules (gr, dark blue), microtubules (light purple) and inner membranes including ER (er, light blue). Source: [191] **(b)** Scale model of the *O. tauri* spindle in the context of the chromatin, shown in longitudinal cross-section. A small bundle of incomplete microtubules (green) resides at each pole, and up to one long, incomplete microtubule extends deep into the nucleus from each pole. Heterochromatin (blue) forms a torus-like structure with a central channel called the 'spindle tunnel'. Individual chromosomes could not be resolved. The nuclear envelope (not shown) has openings at both spindle poles. Source: [194] **(c)** Hypothetical model of polymer melt chromatin in a mitotic cell, viewed along the spindle axis. Canonical nucleosomes (light blue spheres) and centromeric nucleosomes from nonhomologous chromosomes (multicoloured spheres) are positioned around the spindle. Kinetochore protein complexes (lilac rods) connect the centromere to the spindle microtubules (green rings). Source: [195]

2.2.3 *Micromonas*

In 1951, an autotrophic flagellate less than 2μm in size is found in the coastal waters outside Plymouth, England and is named *Chromulina pusilla* Butcher [212], the first ever picoplanktonic species to be described. Afterwards, the pigments and fine structures point towards a position in or near the Chlorophyceae instead of Chrysophyceae (to which the genus *Chromulina* belongs) [213], requiring a change in taxonomic position and a new name: *Micromonas pusilla* [214]. Initial studies focussed primarily on its flaggelar movement [215] and virus infection [216–218]. *M. pusilla* was also shown to be the dominant member (avg. 22%) of the phytoplankton community all year round [219], and is dispersed globally [167, 219, 220].

Like *Ostreococcus*, the genus *Micromonas* contains several genetic lineages or clades. At first three clades were detected using molecular probes [168] and this was later expanded to 5 clades [221].

The genome sequence

In 2009, two isolates (RCC299 and CCMP1545) were sequenced [221]. The genomes were larger (20.9 and 21.9 Mb) than *Ostreococcus tauri* and contained more genes (10,056 and 10,575) but remain very compacted with small intergenics and high coding density. The heterogeneity found in *Ostreococcus* resurfaced in *Micromonas*, resulting in two regions with no collinearity, different codon usage, and a diverging GC content. Intronic repeat elements were detected in one isolate (CCMP1545) that were entirely absent in the other. Their high similarity and copy number reminds of transposable elements, resulting in the name 'Introner Element'. 9,904 IEs were detected and classified into four subfamilies: IE1 to IE4.

2.2.4 *Bathycoccus*

The genus *Bathycoccus* and its first species *Bathycoccus prasinos* are first isolated and described in 1990 [165]. It's a nonmotile ellipsoidal organism with one mitochondrion, one chloroplast, one Golgi complex, and an intricate layer of scales covering the cell surface. The latter are almost circular and have a spider web-like pattern of radiating and concentric ribs. Many features are characteristic for prasinophytes, and the spider web-like pattern is typical of the order Mamiellales – although some members have lost these scales – placing *Bathycoccus* next to the other genera *Ostreococcus*, *Micromonas*, *Mantoniella* and *Mamiella*.

Until recently, all *Bathycoccus* strains were grouped into the same clade [1], but further analysis confirmed the existence of two ecotypes associated with oceanic (more nutrient poor) or mesotrophic (with a moderate amount of dissolved nutrients present) environments [222].

2.2.5 Seagrasses & *Zostera*

After the transition from water to land, several flowering plants (Angiosperms) returned to the aquatic environment ('back to the sea'). One group is called mangroves; trees and shrubs that thrive in partially submerged conditions. The second group belongs to the monocot order Alismatales and its members are dubbed seagrasses. These monocotyledonous plants grow in the sea bed sediment in the coastal waters of most continents. Seagrasses are amongst the most productive ecosystems on a global scale. They provide shelter for a range of marine organisms such as fish and invertebrates and are a major food source for grazing animals. Their rhizomes and roots help to anchor the seabed, preventing erosion of coastal shores.

The recolonisation of the sea did not happen instantaneously. Instead, an intermediate freshwater environment was first colonized from where a transition to marine waters could occur, probably by salt-tolerant species that acquired the necessary adaptations [223]. This evolution seemed to have occurred three times within the Alismatales [224] and resulted in four seagrass families: Zosteraceae (marine), Cymodoceaceae (marine), Posidoniaceae (marine) and Hydrocharitaceae (marine & freshwater) *(Figure 2.7)*. Members belonging to the Ruppiaceae family (e.g. *Ruppia maritima*, a salt-tolerant freshwater species) are sometimes labelled as a seagrass, but are generally excluded because they rarely fully penetrate marine environments.

Seagrasses share a common architecture. They are composed out of several genetically identical units, each consisting of a root system, several leaves and a part the rhizome (i.e. stems extending horizontally below the sediment surface). All clones are linked to the same rhizome, forming a cluster, or 'patch' of clonal shoots. This rhizome allows the seagrass to reproduce asexually simply by extending the rhizome and generating more units. Sexual reproduction is also possible, and all but one marine angiosperm species (Enhalus acoroides [225]) are water-pollinated ('hydrophile'). For the pollen to find their way to the stigma in the water, several adaptations were made e.g. changes in shape and form of the pollen and the loss of the outer wall layer (exine) [226].

More than 100 years ago, the first seagrass studies were being performed on members of the *Zostera* genus – commonly called 'eelgrass' – and more specifically the species *Zostera marina* [227, 228]. The genus *Zostera* is the most widespread seagrass genus in the world [229] and has between 15 and 17 recognized species depending on the source. While most of the seagrasses are dioecious (male and female flowers on different plants), *Zostera marina* is monoecious.

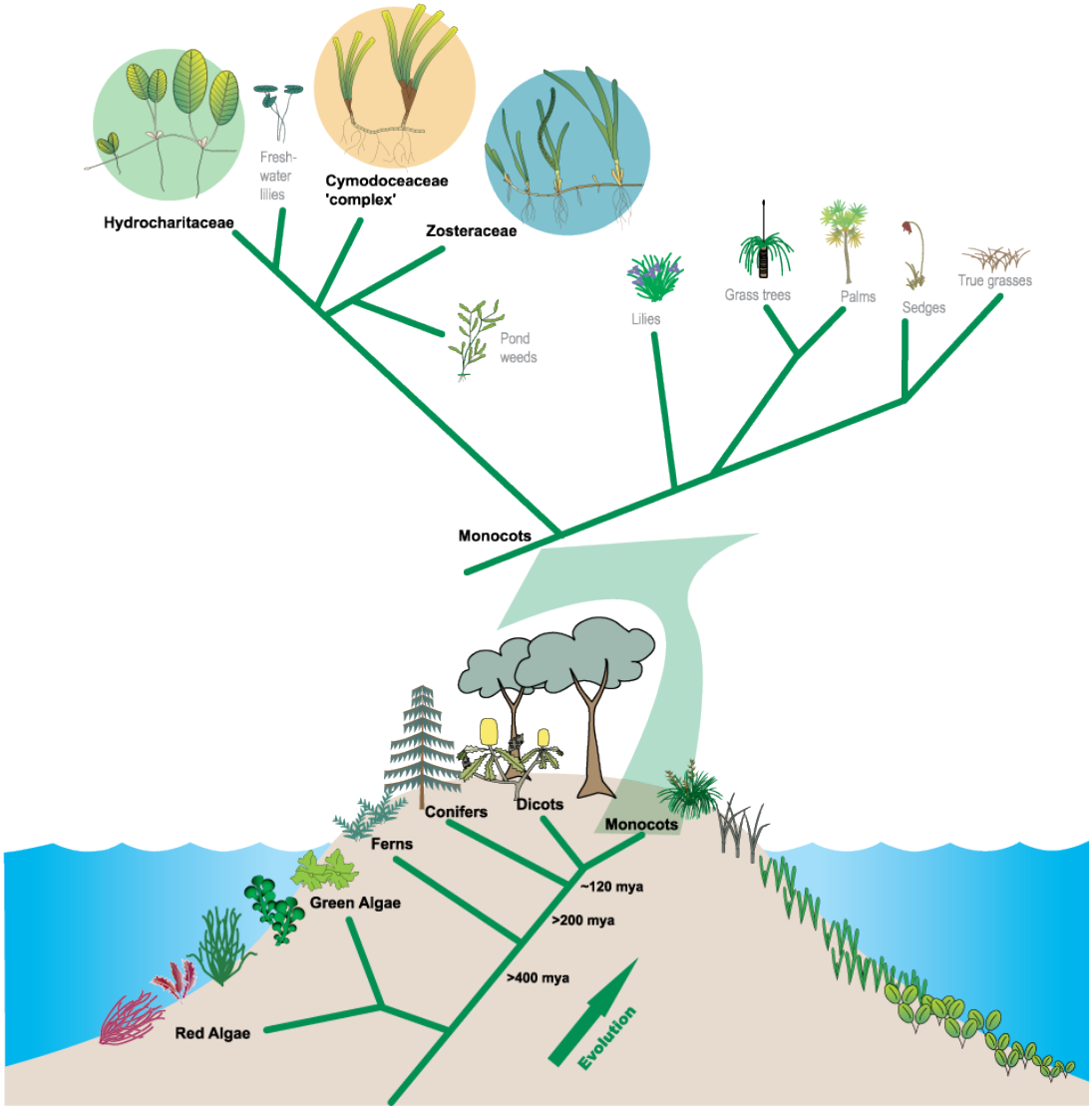


Figure 2.7. Conceptual diagram illustrating the evolution of seagrass species. After the transition from water to land, seed plants eventually arise within the green lineage. Within these vascular land plants we find the Angiosperms, a clade divided into two major groups: monocots and dicots. The monocot lineage is displayed in more detail. The Posidoniaceae and Ruppiaceae families are included within the 'Cymodoceaceae complex' [224]. Image credited to Catherine Collier, James Cook University (2004), IAN/UMCES Symbol and Image Libraries.

Eelgrass is regarded as an important indicator species due to its sensitivity to eutrophication. When an overload of nutrients is added to an ecosystem (e.g. through sewage or detergents), phytoplankton communities will quickly rise in numbers (e.g. algal blooms). The phytoplankton will cause shading for *Zostera*, while this plant already requires more light than any other aquatic plant [230]. Additionally, the phytoplankton will use up all the oxygen within the ecosystem and increase sedimentation, events that will add more pressure to seagrass communities. Global warming adds to this complexity, with heat waves and changing salinity levels likely to inflict even more losses in the future [231–233].

2.3 Introns & Splicing

> "... after the original transcription of DNA into a long RNA, regions of this RNA are spliced out: some stretches excised and the remaining portions fused together by an as yet undefined enzymatic process. The exons, regions of the DNA that will be expressed in mature message, are separated from each other by introns, regions of DNA that lie within the genetic element but whose transcripts will be spliced out of the message" (Walter Gilbert, Nobel Prize in Chemistry, 1980, Nobel lecture on DNA sequencing and gene structure)

Whenever a gene is transcribed, the newly-synthesised RNA transcript is processed – 5' cap addition, splicing, RNA editing and 3' polyadenylation – before it can be translated into a protein. Splicing is the process whereby introns are removed from the transcript, and exons are joined together. Several splicing pathways exist in nature, each related to one or more groups of introns: self-splicing (group I, II and III introns), tRNA splicing (tRNA introns), and the spliceosomal pathway (U2 and U12 introns).

Group I and II introns are self-splicing introns; large catalytic RNAs – called RNA enzymes of Ribozymes – who excise themselves from their RNA precursors without any assistance from protein complexes. Group I introns are widespread in bacterial, phage, viral and organellar genomes as well as nuclear ribosomal DNA (rDNA) genes in fungi, plants and algae [234, 235]. They employ an external guanosine nucleotide as a cofactor in a two-step splicing pathway. Group II introns can be found in organellar, bacterial and archaeal genomes, but also show up in nuclear genomes [236]. The mechanism involved is identical to that of group I introns with the exception that an adenine residue within the intron itself acts as the cofactor. Although some group II introns are able to self-splice *in vitro*, they require some assistance for *in vivo* efficient splicing, presumably to help stabilize and fold the intron RNA into the catalytically-active structure. The assistance is found within the group II intron on an open reading frame that encodes for a Reverse Transcriptase-like protein that also acts as a maturase (splicing factor) [237, 238]. A third group (group III introns) only contains a limited amount of representatives and it shares many features with group II introns, although the intron size is much smaller and splice sites are less conserved [239]. Little is known about their splicing and they are often excluded from intron reviews.

A second splicing pathway is specific for tRNA genes. In bacteria, the tRNA introns are group II introns (see previous paragraph), of which the splicing is RNA-mediated. In archaeal and eukaryotic nuclear genomes however, endonucleases and ligases are needed to produce complete tRNA molecules [240]. The splicing process is highly influenced by the tertiary structure of the tRNA and only a few key nucleotide positions play a role [241].

The last splicing pathway employs the services of the spliceosome, a highly dynamic and complex assembly of RNA and protein components [242–245]. Spliceosomal introns, a hallmark of eukaryotes, are subdivided into two types [246]. The majority (~99%) belong to the U2 type and are excised by the 'major spliceosome'. A minority of the introns contains different splice signals (U12 type) and requires the help of a 'minor spliceosome'. While some spliceosomal components are different between both groups, in both cases the splicing process brings together both ends of the intron, leading to the bulging out of an adenine nucleotide to initiate the splicing process.

The group II self-splicing and spliceosomal splicing mechanism occur in two consecutive transesterification reactions that result in the removal of the intron *(Figure 2.8)*. First, the 2' hydroxyl of the adenine nucleotide attacks the 5' splice site phosphate bond, inducing a break in the RNA while simultaneously creating an intron loop structure, the lariat. Afterwards, the released 3' hydroxyl at the donor splice site will attack the phosphate at the acceptor splice site, thereby linking both exons and releasing the intron lariat (which will be degraded later on).

2.3.1 Intron origin

When introns were first discovered in 1977 [247, 248] and a first intron loss was documented in 1980 [249], scientists were divided trying to explain their presence and evolution *(Table 2.2)*. One view, the Introns-Early (InE) theory, stipulates that introns are ancient and were already present in the Last Universal Common Ancestor

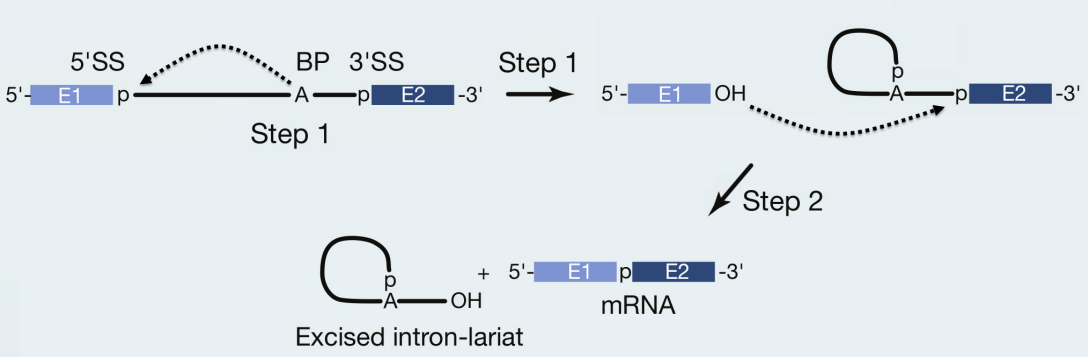


Figure 2.8. Transesterification reactions during the splicing process. For efficient splicing, group II introns require the help of maturases, while spliceosomal introns employ the help of the spliceosome. 5' SS: donor splice site; 3' SS: acceptor splice site; BP: branch-point; E1: exon1; E2: exon2. Adopted from [243]

(LUCA) of prokaryotes and eukaryotes *(Figure 2.9a)*. Bacteria were said to be streamlined and thus devoid of all introns. The Introns-Early theory ties in with the 'exon theory of genes', which states that early genes were created through the intron-mediated shuffling of exons [250, 251], whereby each exon coded for a module or protein domain, and introns are actually ancient intervening/intergenic regions. This would then explain the over-representation of phase-0 introns i.e. introns lying in between two codons. During evolution, intron loss has been the most frequently observed event by far, indicating that the LUCA did contain some spliceosomal introns [252, 253]. The Introns-Late (InL) theory proposes that introns emerged later, after the divergence of eukaryotes and prokaryotes *(Figure 2.9b)*. Spliceosomal introns are the crux of the InE-InL debate, while the other intron groups are generally only considered in relation to the evolution towards said spliceosomal introns (see next section).

Introns-Early	**Issue**	**Introns-Late**
genome streamlining: ancient intergenic regions were lost	**no spliceosomal introns in Bacteria**	introns arose after the divergence of prokaryotes and eukaryotes
no explanation: exon theory of genes falls short	**no correlation between exons and protein domains**	introns inserted into undivided genes (no relation to modules or domains)
phase-0 introns represent ancient intergenic regions	**over-representation of phase-0 introns**	introns are inserted into different sequence patterns at different frequencies, resulting in a non-random phase distribution
introns arose very early	**conservation of intron positions in distant eukaryotic taxons**	the early eukaryotic ancestor contained many introns

Table 2.2. Introns-Early versus Introns-Late. Sources:[254, 255]

Both models still do not explain the enormous diversity in intron numbers between eukaryotic species, with intron-rich and intron-poor species interspersed in the eukaryotic tree. This could imply either an intron-rich ancestor with subsequent lineage-specific losses, or an intron-poor ancestor with subsequent lineage-specific gains. Currently, the consensus view states that the early eukaryotic ancestor contained relatively large number of introns. During intron evolution, intron gain is considered a rare event, while intron loss is more common, though more variable and lineage-specific [257, 258].

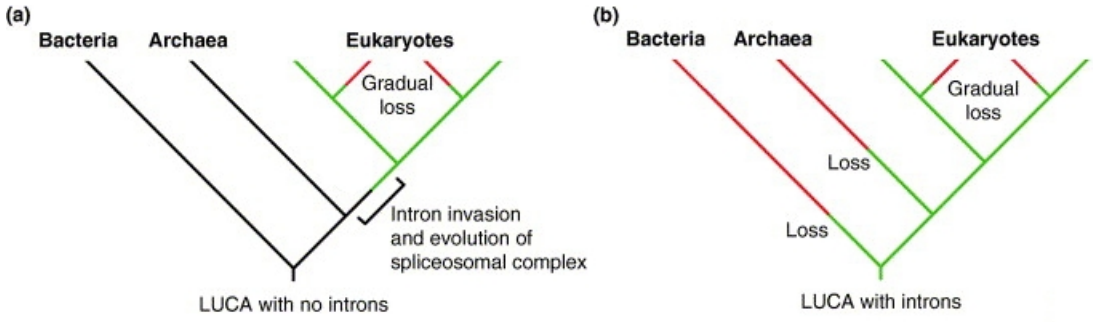


Figure 2.9. The origin of introns according to the Introns-Late (a) and Introns-Early (b) theory. Green branches indicate lineages containing introns, while red branches indicate loss of introns. Black branches denote pre-intron stages. LUCA: Last Universal Common Ancestor. Source: [256]

2.3.2 Evolution towards spliceosomal introns

The excision of group II introns will result in the formation of a lariat [259], and the intron-encoded maturase is reminiscent to eukaryotic splicing factors. Even the actual splicing mechanism (two successive transesterification reactions) is identical between group II and spliceosomal introns. Additionally, the active site within the spliceosome shares remarkable similarities to that of group II introns [260], as do the small nuclear RNAs (snRNAs) [261]. Further analysis revealed a trans-splicing capability in both group I and group II introns [262, 263] and the existence of an organellar spliceosome to splice out specific group II introns [262, 264]. Such structural and mechanistic similarities led many to develop the hypothesis that group II introns are the progenitors of both the spliceosome and spliceosomal introns, and have thus been highly important in shaping the genome during eukaryotic evolution [236, 265, 266].

Even outside the organelle, group II introns are able to splice accurately from nuclear transcripts [267]. However, the subsequent transcripts are subject to nonsense-mediated mRNA decay (NMD), and are poorly translated. The overall outcome implicates that the group II intron should either be lost, or that it could evolve and possibly require the help of proteins for effective splicing and further processing. The latter ties in nicely with the ancestry hypothesis described in the previous paragraph. How and when this evolution occurred is still highly debatable [268] *(Figure 2.10)*.

2.3.3 Intron functionality

Why did introns and splicing arise? It's quite hard to fathom why an organism would bring upon itself the burden of 1) transcribing 'unnecessary' regions, 2) splice the transcripts via the spliceosome, one of the cells largest complexes requiring vast amounts of energy and resources to construct, and 3) break down the remaining lariat products. If even one of these steps would be interrupted, the cell would be in grave peril. While it is possible that the cell adapted to the presence of introns and tried to cope with the situation, it is hardly imaginable the introns would not serve a purpose. One of the earliest intron functions was 'intron-mediated enhancement', a boost in expression levels of genes due to the presence of an intron. The mechanism through which introns boost transcription initiation involves the presence of cis-regulatory elements – or even entire alternative promotors – within the intron sequence [270], or by creating a favourable local chromatin structure [271]. While the previous example demonstrates intron functionality based on the genomic intron sequence (no transcription/splicing/... has yet occurred), an intron has many functional forms.

When an intron is actively transcribed or spliced out, it plays a huge role in transcript regulation. During transcription, long introns induce time delays between the actual gene activation and the appearance of the protein product. During splicing, different splicing factors associate or interact with transcription factors, thereby influencing transcription initiation, elongation and termination [270]. An example is the U1 snRNA-mediated pre-initiation complex, a stage during the spliceosome assembly that focuses on the donor splice site. This complex associates with TFIIH, TFIID and TFIIB, general transcription factors that are important for transcription initiation. Additionally, the spliceosome facilitates the recruitment of export factors important for

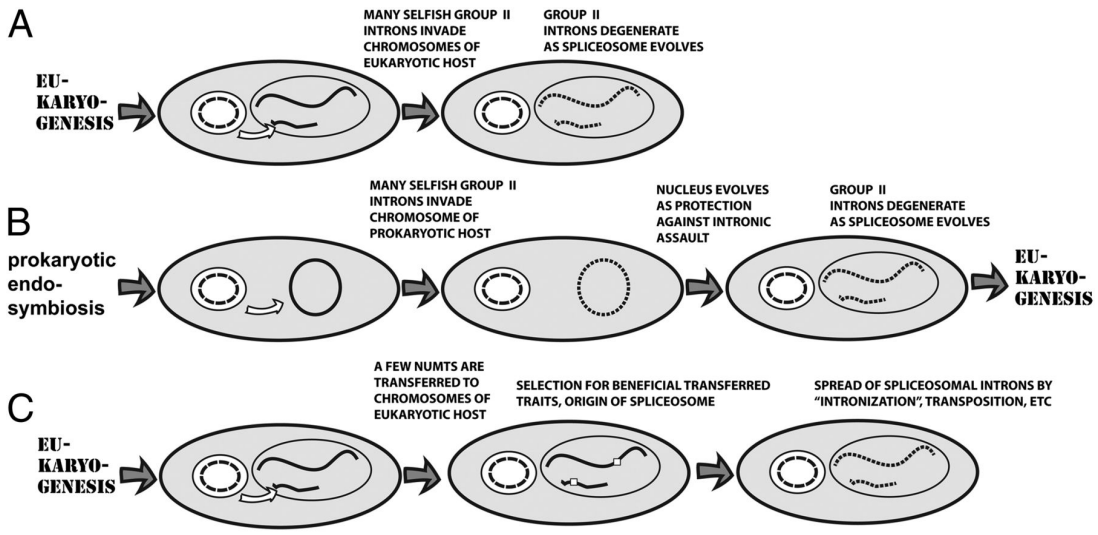


Figure 2.10. Schemes for the evolution of spliceosomal introns. (a) In a relatively advanced eukaryotic cell, selfish group II introns from newly established mitochondria invade the nuclear genome. They proliferate and subsequently degenerate into spliceosomal introns [269]. **(b)** Selfish group II introns from an endosymbiotic bacterium invade the genome of a host Archaean and harmfully proliferate. This proliferation forces the evolutionary separation of nucleus and cytoplasm and the evolution of spliceosomes, resulting in the rise of eukaryotic cells [265]. **(c)** In a relatively advanced eukaryotic cell, organellar lysis releases pieces of mitochondrial DNA that are transferred to the nuclear genome – numts – and contain group II introns. This nuclear copy of the mitochondrial gene can confer some advantage, resulting in positive selection. Source: [268]

transporting the mRNA out of the nucleus. After the splicing process has terminated, RNA genes embedded within the excised introns are expressed. This allows genes to auto-regulate their own expression, usually through micro-RNAs, short non-coding RNA (ncRNA) sequences that bind mRNA target sites and redirect them for degradation.

2.3.4 GC content & splice site recognition

Spliceosomal introns exhibit innate properties that allow the spliceosome to separate pseudo splice sites from true splice sites. Sequence motifs such as the donor, acceptor and branch point interact with the spliceosome to define exon/intron boundaries and initiate splicing. The GC content – i.e. the percentage of guanine/cytosine bases on a DNA sequence – also plays a major role in determining the exon/intron boundaries, and even influences intron length [272]. Studies have established that introns in GC-poor regions are flanked by exons with remarkably higher GC content. This drop in GC content acts as a flag marking the exon/intron boundary. This flag could compensate for very lengthy intron sequences, and allow the exon to stand out and be recognised by the spliceosome. In GC-rich regions however, introns have roughly the same GC content as their flanking exons.

We already mentioned the influence of GC content on intron length. Intron length is very dependent on the species and the lineage. For plants, the distribution peaks at ~100 nucleotides, with a right tail containing the low-abundance long introns [273]. The intron length influences the recognition of the splice site boundaries: long introns would favour a mechanism that identifies (shorter) exons (i.e. exon definition), while short introns would favor intron definition [274].

2.3.5 Intron mobility

Introns can act as mobile genetic elements in a variety of mobility reactions. Group I introns often contain an open reading frame for a homing endonuclease, a protein that drives the mobility and persistence of its own reading frame. This endonuclease creates a double-strand DNA break within the allele of the host gene

that does not contain the group I intron. Using homology-driven DNA repair, the uninterrupted allele will now contain the group I intron [235, 275, 276]. Intron homing is also possible at the RNA level by reverse splicing [275]. In this mechanism, an excised intron recognises an insertion site within a transcript and integrates itself followed by reverse transcription and genomic integration (through recombination).

Group II introns are involved in two types of mobility reactions: retrohoming, in which they insert at high frequency into specific sites, and retrotransposition, in which they insert at low frequency into non-specific sites. Retrohoming targets DNA directly with help from the intron-encoded protein (IEP), which acts as a Reverse Transcriptase, maturase, and endonuclease *(Figure 2.11)*. If there is no endonuclease activity, a nascent strand of the replication fork is used for priming reverse transcription (which could be the predominant mechanism in group II mobility pathways) [277]. Because it uses the same base-pairing interactions for both DNA integration and RNA splicing, the group II intron ensures it is able to insert only at those target sites from which it can subsequently excise. Group II introns can also insert into non-cognate (~non-allelic) sequences that resemble the homing site. This process occurs via an RNA intermediate (as described in the previous paragraph for group I introns) at lower frequencies and it is known as ectopic transposition, or retrotransposition [278–280].

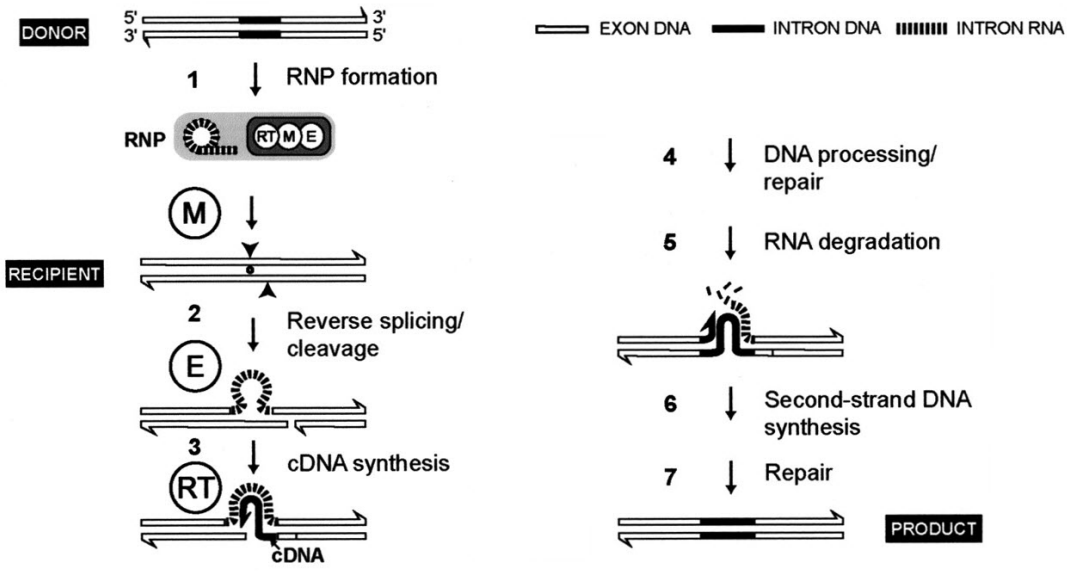


Figure 2.11. Group II retrohoming pathway for the Ll.LtrB intron in *L. lactis* and *E. coli*. (1) After transcription and splicing, the excised intron lariat and the translated IEP will form a ribonucleoprotein (RNP) particle having Reverse Transcriptase (RT), maturase (M) and DNA endonuclease (E) activities. **(2)** The RNP will recognise a DNA target site. The interactions formed allow separation of the DNA strands and integration – by reverse splicing – of the group II intron in one strand. The other strand is cleaved. **(3)** The Reverse Transcriptase will subsequently start complementary DNA (cDNA) formation using the cleaved 3' end as primer (target-primed reverse transcription). **(4-7)** Extended cDNA synthesis, RNA degradation, second-strand DNA synthesis, and repair. The order of steps 4-7 is unknown. Source: [281]

2.3.6 Intron gain mechanisms

Several mechanistic origins have been proposed for intron gains [282, 283], mainly based on indirect evidence only. *In vivo* demonstrations have only been established for the 'Tandem genomic Duplication' mechanism [283, Table 1].

Previously mentioned, intron transposition involves the reverse splicing of a spliced-out intron into a transcript (either allelic or non-allelic). After reverse transcription, the intron-containing transcript will be integrated into the genomic sequences via complete or partial recombination. Because this mechanism does not require the generation of splicing signals – they are already present in the transposed intron – it is the favoured intron gain mechanism of many researchers, even when considering the pitfalls (see next paragraphs).

The concept of reverse splicing (and the subsequent reverse transcription and recombination) is very important in the light of intron mobility and intron gain. In 2008, a study unveiled that both catalytic steps of nuclear precursor messenger RNA (pre-mRNA) splicing could be reversed [284], albeit by using a mutant that fails to release the mRNA from the spliceosome. The actual mechanism that allows the spliced intron – and the spliceosome with whom it is still affiliated – to recognize a transcript and initiate revere splicing, is not known. Additionally, no reverse spliced introns have ever been found in EST or cDNA sequences, leading many to believe that reverse splicing is an extremely rare process [282]. Nevertheless it is essential to many, often theoretical, intron mobility pathways.

Reverse Transcriptase (RT) is also a key player in many pathways. Reconstruction of intron gain and loss rates in 19 different eukaryote species indicates RT plays an important role in the efficient removal of introns, and little to no role in intron gain [285]. The only difference between the RT-mediated intron loss mechanism and the intron transposition pathway is the reverse splicing step. This means that the difference between intron loss and gain depends solely on the rate of reverse splicing, which only occurs at low frequency. The balanced rates of intron gain and loss in specific lineages thus challenge the traditional intron transposition model. Possible explanations are different rates of genomic recombination for intron gain and intron loss, or the presence of sequence signals within some introns that favour their reverse transcription and/or cDNA re-importation into the nucleus.

Two alternative mechanisms were also proposed that allow introns to propagate to novel sites [282], the first of which is called 'spliceosomal retrohoming' *(Figure 2.12a)*. Much like the retrohoming pathway of group II introns, it involves reverse splicing directly in the genome instead of using a transcript intermediate *(Figure 2.11)*. The main selling point for this model is the absence of the transcript phase, whereby a novel intron reverse splices into a mature messenger RNAs, running the risk of potential degradation by nonsense-mediated mRNA decay. Additionally, the model combines reverse splicing and genomic integration into one single simultaneous action, whereas the intron transposition model requires two separate steps. This implies that even a small rate of this process could yield a high yield of intron gain events. The second alternative involves template switching during reverse transcription *(Figure 2.12b)*. During reverse transcription of a transcript, the nascent cDNA strand can dissociate from the template strand. Using its 3' terminus, it can re-associate with another complementary RNA region and continue the reverse transcription process, creating a chimeric cDNA molecule [282]. In this case, template switching during reverse transcription of a genic RNA can create a chimeric sequence containing an intron. Genomic recombination produces a non-canonical x-shaped structure that can be resolved in several ways, potentially leading to a viable intron structure.

Additionally, instead of focusing on the proliferation of already existing spliceosomal introns, the mechanisms detailed above could target classes of repeat-like introns. After such repeat introns experience bursts of proliferation, they slowly degenerate (losing their repeat character) and start to resemble regular spliceosomal introns (RSIs). This mechanism can – over time – provide gains of spliceosomal introns in specific lineages [286–288].

2.3.7 Intron position conservation

We already established that an early eukaryotic ancestor contained relatively large amounts of introns. This theory explains why many intron positions are conserved across the different eukaryotic lineages [289, 290]. We also have to account for 'intron sliding', a phenomenon that allows for positional variation through the relocation of intron-exon boundaries over short distances (1-15bp) [291]. While alignment artefacts account for a great portion of the 'sliding', results strongly suggest that one-base-pair sliding is a real evolutionary phenomenon occurring in <5% of all introns [292].

Most intron positions however are not conserved, and are due to lineage-specific intron gains. Additionally, some positions that seem conserved at first sight, are actually not. Intron gains can occur in different lineages at the exact same position, which resembles conserved intron positions [293, 294]. Such parallel intron gains are rare and have only been extensively documented in the microcrustacean *Daphnia pulex* [295, 296] and fungi [297]

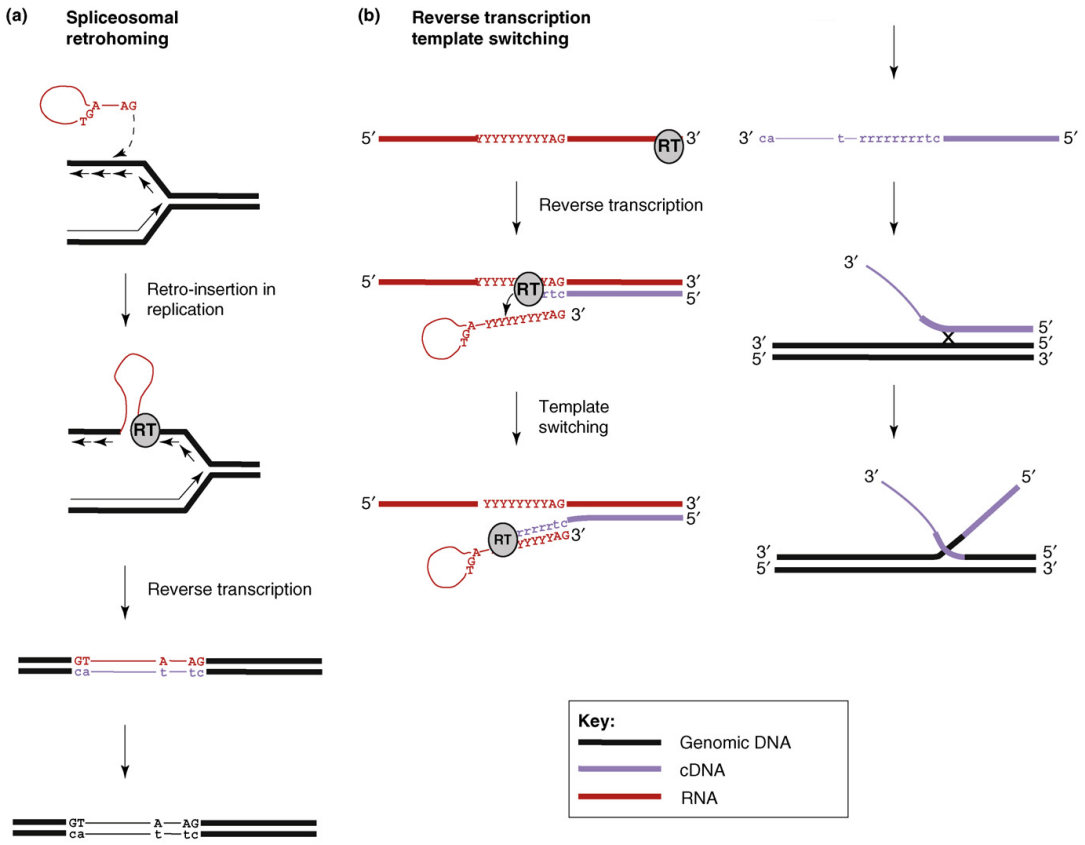


Figure 2.12. Alternative intron gain mechanisms. **(a)** Spliceosomal retrohoming. The intron lariat reverse splices directly into the genome. **(b)** Template switching during reverse transcription creates a chimeric cDNA. Source: [282]

CHAPTER 3

GENE FUNCTIONALITIES AND GENOME STRUCTURE IN *BATHYCOCCUS PRASINOS* REFLECT CELLULAR SPECIALIZATIONS AT THE BASE OF THE GREEN LINEAGE

*Hervé Moreau, **Bram Verhelst**, Arnoud Couloux, Evelyne Derelle, Stephane Rombouts, Nigel Grimsley, Michiel Van Bel, Julie Poulain, Michaël Katinka, Martin F Hohmann-Marriott, Gwenaël Piganeau, Pierre Rouzé, Corinne Da Silva, Patrick Wincker, Yves Van de Peer and Klaas Vandepoele*

Abstract

Bathycoccus prasinos is an extremely small cosmopolitan marine green alga whose cells are covered with intricate spider's web patterned scales that develop within the Golgi cisternae before their transport to the cell surface. The objective of this work is to sequence and analyse its genome, and to present a comparative analysis with other known genomes of the green lineage. Its small genome of 15 Mb consists of 19 chromosomes and lacks transposons. Although 70% of all *B. prasinos* genes share similarities with other Viridiplantae genes, up to 428 genes were probably acquired by horizontal gene transfer, mainly from other eukaryotes. Two chromosomes, one big and one small, are atypical, an unusual synapomorphic feature within the Mamiellales. Genes on these atypical outlier chromosomes show lower GC content and a significant fraction of putative horizontal gene transfer genes. Whereas the small outlier chromosome lacks colinearity with other Mamiellales and contains many unknown genes without homologs in other species, the big outlier shows a higher intron content, increased expression levels and a unique clustering pattern of housekeeping functionalities. Four gene families are highly expanded in *B. prasinos*, including sialyltransferases, sialidases, ankyrin repeats and zinc ion-binding genes, and we hypothesize that these genes are associated with the process of scale biogenesis. The minimal genomes of the Mamiellophyceae provide a baseline for evolutionary and functional analyses of metabolic processes in green plants.

Contributions

- Structural and functional genome annotation (including repeat detection)
- Manual gene curation
- Set-up and maintenance of the *Bathycoccus*-section on the ORCAE platform
- BOC/SOC analysis (definition, expression, introns, GC%, comparison with other microalgae)
- Figures *(Figures 3.3 and 3.4 and Supplementary Figures 3.2 to 3.4)*
- Tables *(Supplementary Tables 3.1 to 3.3)*
- Writing manuscript segments (in respect to the topics mentioned above)

3.1 Introduction

Marine phytoplankton is responsible for about half of the photosynthetic activity on the planet [298], the second half being carried out by terrestrial plants. Two major traits differentiate these two classes of organisms. First, phytoplankton is essentially composed of unicellular organisms that have a high turnover; whereas terrestrial plants are renewed, on average, once every 9 years, the global phytoplankton population is replaced approximately every week [298]. Second, while photosynthesis is confined to specific organs of plants, often only a minor component of the plant biomass, in phytoplankton, photosynthesis essentially takes place in each cell. Phytoplankton populations are thus highly dynamic and may be able to adapt rapidly to changing environments. Even so, a global decline of photosynthetic micro-organisms over the past century has recently been reported [299], motivating research aimed at better understanding the global diversity of phytoplankton and how these species adapt to changing marine environment.

Phytoplankton is usually pragmatically classified according to size, from pico- (below 3 μm), nano- (3 to 8 μm) to micro-algae (above 5 to 8 μm), although these categories have no evolutionary significance. The eukaryotic fraction of picophytoplankton accounts for a modest part of the oceanic biomass, but nevertheless contributes an important part to primary production in many oceanic waters [300, 301]. Among these picoeukaryotes, environmental diversity studies based on ribosomal gene sequences showed that small green algae, and notably the three genera *Bathycoccus*, *Micromonas* and *Ostreococcus*, are distributed worldwide and are numerically important in coastal areas. These three genera are characterized by their small size (1 to 2 μm), their rudimentary cellular organization (one mitochondrion and one chloroplast) and their small genomes (from 13 to 22 Mb). *Micromonas* [212] is a naked cell with one long flagellum whereas the two other genera are non-motile. *Ostreococcus* [170, 171] is naked whereas *Bathycoccus* [165] is covered with scales. The complete genome sequences of two *Micromonas* [221], two *Ostreococcus* [181, 182] and a low-light adapted strain of *Ostreococcus* – strain RCC809, available on the Joint Genome Institute (JGI) web site – have been analysed. The three genera belong to the order Mamiellales, in the class Mamiellophyceae [160, 166], a monophyletic group in the phylum Chlorophyta. The ancestors of these micro-organisms emerged at the base of the green lineage and knowledge about them provides a baseline for exploring the evolution of this lineage, which also gave rise to terrestrial plants. Given their small cellular and genome sizes, they may reveal the 'bare limits' of life as a free-living photosynthetic eukaryotes, thus presenting a simple organization with very little non-coding sequences [302].

Here we report the analysis of the genome of one Mediterranean strain belonging to the genus *Bathycoccus* and its comparison with Mamiellales and other green algae, allowing a survey of the genome organization at the base of the green lineage. Although *Bathycoccus* was initially isolated from deep water (100 meters) [165], it has been frequently reported in various marine environments and seems an important component of the picoeukaryote compartment [303–306]. The availability of this genome, coupled to the development of new sequencing possibilities for metagenomes [307, 308] from various marine environments, opens the way for comparative studies and to a better understanding of the adaptations of this organism to its environment(s).

3.2 Results and discussion

3.2.1 Characterization and phylogenetic position of the *Bathycoccus prasinos* RCC1105 strain

We isolated the *Bathycoccus prasinos* strain RCC1105 from a seawater sample from Banyuls' bay collected in January 2006. Contrary to the type strain described as *Bathycoccus prasinos* [165], which was isolated at a depth of 100 meters, RCC1105 was isolated from surface water (5 m). The strain RCC1105 has a typical *Bathycoccus* morphology with scales covering the cell *(Figure 3.1)* and we confirmed its taxonomic affiliation by PCR amplification of its 18S ribosomal gene. The complete genome of RCC1105 revealed two unlinked identical copies of the rDNA genes. Unlike the two previously reported *B. prasinos* isolates [165, 309], these two ribosomal 18S genes were found to harbor an identical 433 bp long group I intron starting at position 551. Apart from this, the nucleotide sequence was strictly identical to the reference strain (GenBank: AY425315, FN562453). Self-splicing group I introns are widespread in nature, and have been recorded in the 18S rDNA of several other protists [234], including some within the green lineage, but, so far, not within the Mamiellales. All

four *Bathycoccus* strains isolated from the Mediterranean bear this intron located exactly at the same splicing site. Phylogenetic analysis based on this small ribosomal subunit and on the Internal Transcribed Spacer (ITS) confirmed that, in contrast to the two other Mamiellales' genera *Micromonas* and *Ostreococcus*, all *Bathycoccus* strains isolated to date comprise only one clade [160, 166]. To confirm the phylogenetic position of *Bathycoccus* within the Mamiellales, we concatenated a set of 154 single-copy genes conserved in 13 species, including plants, and aligned them over 35,431 amino acids to construct a maximum likelihood phylogenetic species tree *(Supplementary Figure 3.1)*. The phylogeny obtained was well-supported and showed that the genus *Bathycoccus* is closer to *Ostreococcus* than to *Micromonas*.

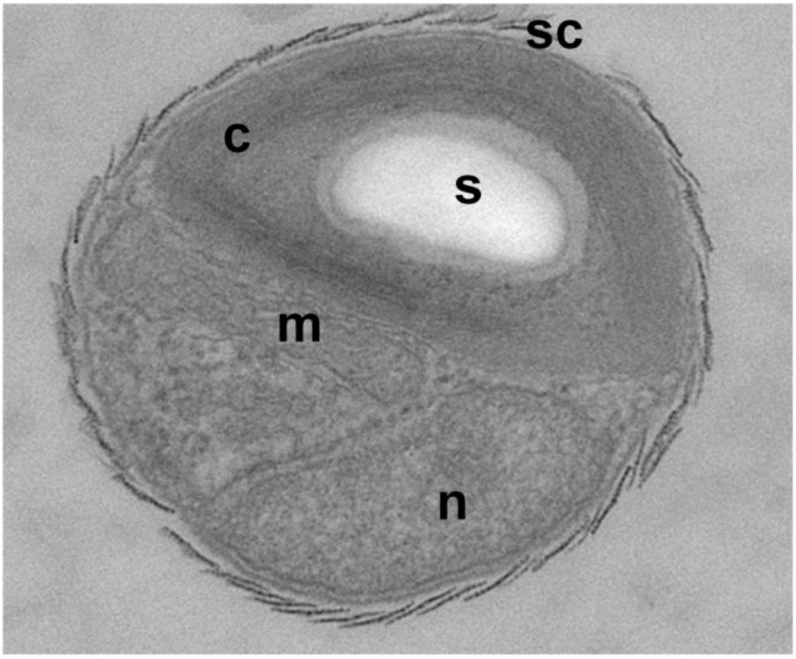


Figure 3.1. Morphology of the *Bathycoccus prasinos* RCC1105 strain. Morphological characterization of the *Bathycoccus* RCC1105 strain: EM picture of an exponentially growing *Bathycoccus* RCC1105 cell. Abbreviations: c, chloroplast; n, nucleus; s, starch granule; sc, scale covering the surface of the cell.

3.2.2 Global characteristics of the *Bathycoccus* genome

The global characteristics of the *Bathycoccus* genome are similar to those observed in other Mamiellales except for its significantly lower GC content [305] *(Table 3.1)*. The global genome size, measured by both Pulsed-Field Gel Electrophoresis *(Figure 3.2a)* and sequencing, is intermediate (15 Mb) between *Ostreococcus* (12 to 13 Mb) and *Micromonas* (21 to 22 Mb), also reflecting an intermediate number of genes *(Table 3.1 and Supplementary Table 3.1)*. Both sequencing and Pulsed-Field Gel Electrophoresis also showed the genome to comprise 19 chromosomes, a number close to that found in other Mamiellales, and in other green algae despite the variation in genome size *(Table 3.1 and Supplementary Table 3.1)*. The 15 Mb genome was sequenced at 22-fold coverage using a whole-genome shotgun sequencing approach, resulting in 126 contigs ranging from 3 to 1,353 kb. According to blast analysis, the 102 smallest of these contigs were bacterial contaminations, whereas the 24 remaining bigger contigs were part of the *Bathycoccus* genome (22 nuclear, 1 chloroplastic and 1 mitochondrial contig). Among the 22 nuclear contigs, six could be joined two by two, giving 19 scaffolds corresponding to 19 chromosomes observed by pulse field electrophoresis. Using intrinsic and extrinsic information, we predicted 7,847 genes in the nuclear genome *(section 3.4.2)*, giving a high gene density similar to other Mamiellales. The validity of a majority of predicted genes was supported either by ESTs (approximately 46%) or by protein similarity (approximately 85%), and approximately 15% of them contain introns. Very few repeat sequences were found and no known or new TEs were detected *(Supplementary Table 3.1)*. The

synteny observed between the chromosomes of *Ostreococcus* and *Bathycoccus (Figure 3.2b)* shows that the genome organization is globally better conserved between these two genera than with the genus *Micromonas*, in agreement with the phylogenetic analysis.

Family	Species	Genome size (Mb)	G+C (%)	Chromosome number	Gene number
Prasinohpyceae	*Bathycoccus* sp. RCC1105	151	48	19	7,847
Prasinohpyceae	*Micromonas* sp. RCC299	20.9	64	17	10,286
Prasinohpyceae	*Micromonas* sp. CCMP1545	21.9	65	19	10,587
Prasinohpyceae	*Ostreococcus lucimarinus* clade A	13.2	60	21	7,805
Prasinohpyceae	*Ostreococcus* sp. RCC809 clade B	13.3	60	20	7,492
Prasinohpyceae	*Ostreococcus tauri* clade C	12.6	59	20	8,116
Trebouxiophyceae	*Chlorella* sp. NC64A	46	67	12	9,791
Chlorophyceae	*Chlamydomonas reinhardtii*	121	64	17	15,143
Chlorophyceae	*Volvox carterii*	138	56	14	14,520

Table 3.1. Nuclear genome characteristics of green algae. Data from [182, 221, 310–312].

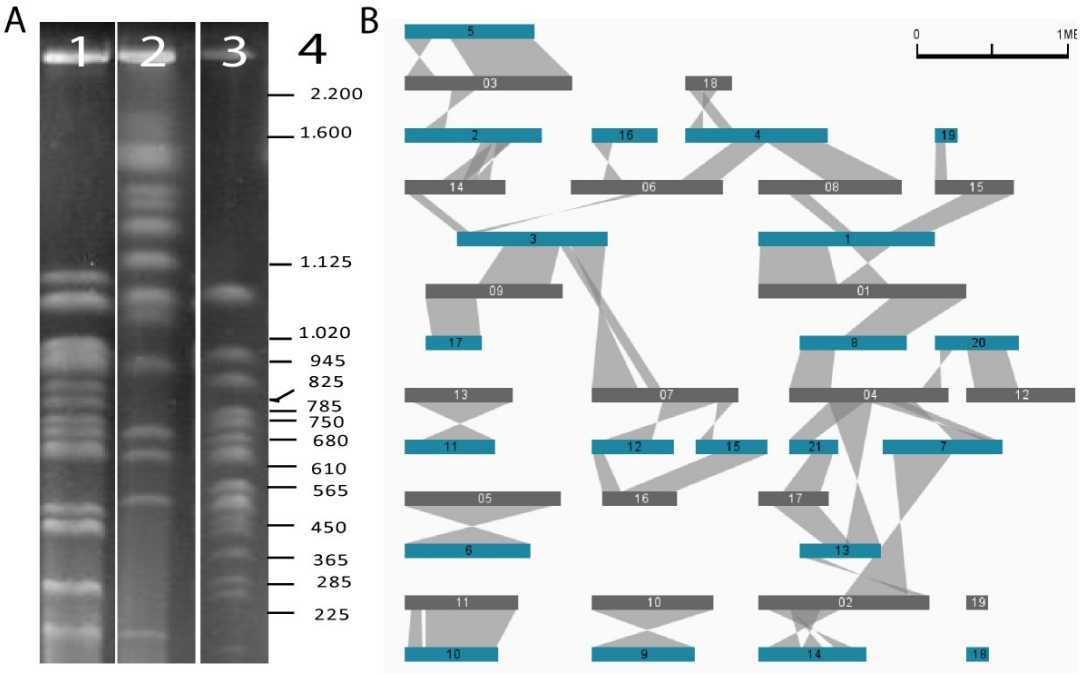


Figure 3.2. Genome organization of the *Bathycoccus prasinos* RCC1105 strain. (a) Pulsed-Field Gel Electrophoresis of the genomes of *Bathycoccus prasinos* RCC1105 (lane 1), *Micromonas pusilla* (lane 2) and *Ostreococcus tauri* (lane 3) DNA fragment length based on the chromosomes of *Saccharomyces cerevisiae* (lane 4). **(b)** Synteny between *Ostreococcus lucimarinus* (blue) and *Bathycoccus prasinos* RCC1105 (grey) genomes.

Based on the annotated gene sets of different land plants and green algae, sequence similarity searches were performed to group homologous genes into families (a family being defined as a set of two or more homologous genes; see Materials and methods). Subsequently, pan and core genome plots were built to quantify the number of shared and unique genes and families between different species. Comparing the set of core genes between different algal groups reveals that the smaller genome sizes of Mamiellales, as well as the

lower number of genes, correspond both with the decrease of the average number of genes per family and with the number of families conserved within a specific clade. For example, whereas the number of gene families shared between all land plants, Chlamydomonales, and Trebouxiophyceae is 2,692, this number drops to 1,959 when including all Mamiellales species. Similarly, based on a set of core gene families conserved in both land plants and algae, the average gene family size is smaller for Mamiellales compared to Trebouxiophyceae or Chlamydomonales (average of 1.63, 1.78 and 1.93 genes per family, respectively). More than 500 gene families were found that were conserved between land plants and green algae but that were lost in all Mamiellales species. These families were enriched for functions related to zinc ion-binding and transport (ten families), UDP-glucosyltransferase activity (six families), vitamin ion binding (eight families) and sucrose and fatty acid metabolism (eight families). Although this pattern suggests a reduction of the functional gene repertoire, we also found more than 400 gene families that are specific to Mamiellales and found in all Mamiellales species. Whereas many of these Mamiellales-specific genes have unknown functions, three families related to drug transport and ten families including genes related to zinc ion binding were found. Although rapid sequence evolution can interfere with the accurate detection of homologs using similarity searches, the observed pattern indicates a high turnover of zinc ion binding-related genes.

3.2.3 Biological role and evolution of the big and small outlier chromosomes in *Bathycoccus* and in the Mamiellales

Despite the low average GC content (48%) of the *Bathycoccus* genome compared to other members of the Mamiellales (over 59%), two outlier chromosomes were found, one 'big' (chromosome 14) and one 'small' (chromosome 19), with lower GC content (42%) compared to the rest of the genome *(Table 3.1, Supplementary Table 3.2, and Supplementary Figure 3.2)*. This kind of organization was previously reported in *Micromonas* and *Ostreococcus* [181, 182, 221, 313] and thus is a characteristic of all Mamiellales that have been sequenced so far. In all species, the atypical genomic features for the Big Outlier Chromosomes (BOCs) are restricted to a sub-region (referred to as BOC1) of the complete chromosome, whereas the whole length of the Small Outlier Chromosome (SOC) shows low GC content *(Supplementary Figure 3.2)*. However, although a BOC region was found for the 'low-light' *Ostreococcus* sp. RCC809 genome, which is available on the Joint Genome Institute website (unpublished), no clear SOC could be identified *(section 3.5.3)*. Whether this observation is biologically correct or the consequence of the applied sequencing approach, read filtering, or genome assembly remains currently unclear. Similar outlier chromosomes have not been found in other green algae such as *Chlamydomonas*, *Volvox* or *Chlorella*. In *Chlorella* low GC chromosome regions were reported [310], but these were, in contrast to those in the Mamiellales, scattered throughout different chromosomes. Outlier chromosomes are highly diverged in terms of gene content. Whereas most *Bathycoccus* chromosomes share, to some extent, a conserved genome organization with the other Mamiellales, both BOC1 (217 annotated genes) and SOC (72 annotated genes) lack colinearity *(Figure 3.3)*, and this pattern is largely conserved between the outliers of the three genera. Many BOC1 genes share orthologs with other Mamiellales while SOC comprises mainly unknown, species-specific genes with few introns (26% of the SOC proteins have Gene Ontology (GO) functional annotation versus 71% for BOC1 genes and 44% for the rest of the genome) *(Figure 3.3)*. Additionally, phylogenetic estimations of the proportions of genes lacking plant orthologs yielded 75% (54/72) for SOC, 16% for BOC1 and 25% for normal chromosomal regions.

3.2.4 The big outlier chromosome in *Bathycoccus*

The size of the *Bathycoccus* BOC is 663,424 bp. Fifty-two and seventy-eight percent of the *Bathycoccus* BOC1 genes having orthologs in other species were also located in the BOC in *Micromonas* and *Ostreococus*, respectively *(Figure 3.4)*. In contrast, the locations of 29 BOC1 single-copy conserved gene markers (that is, genes having orthologs and located in BOC1 in all Mamiellales *Supplementary Table 3.3*) were scattered throughout the genomes in *Chlamydomonas*, *Volvox* and *Chlorella*, revealing that, despite the absence of colinearity, the clustering of the BOC1 genes is conserved and unique to the Mamiellales. These data suggest that BOC1 is a conserved genome property that was present in the last common ancestor of the Mamiellales. Genes located in the BOC1 region are over-represented in basic housekeeping functions like primary metabolism, gene expression, photosynthesis and protein transport *(Supplementary Table 3.3)*. To identify genomic features that are

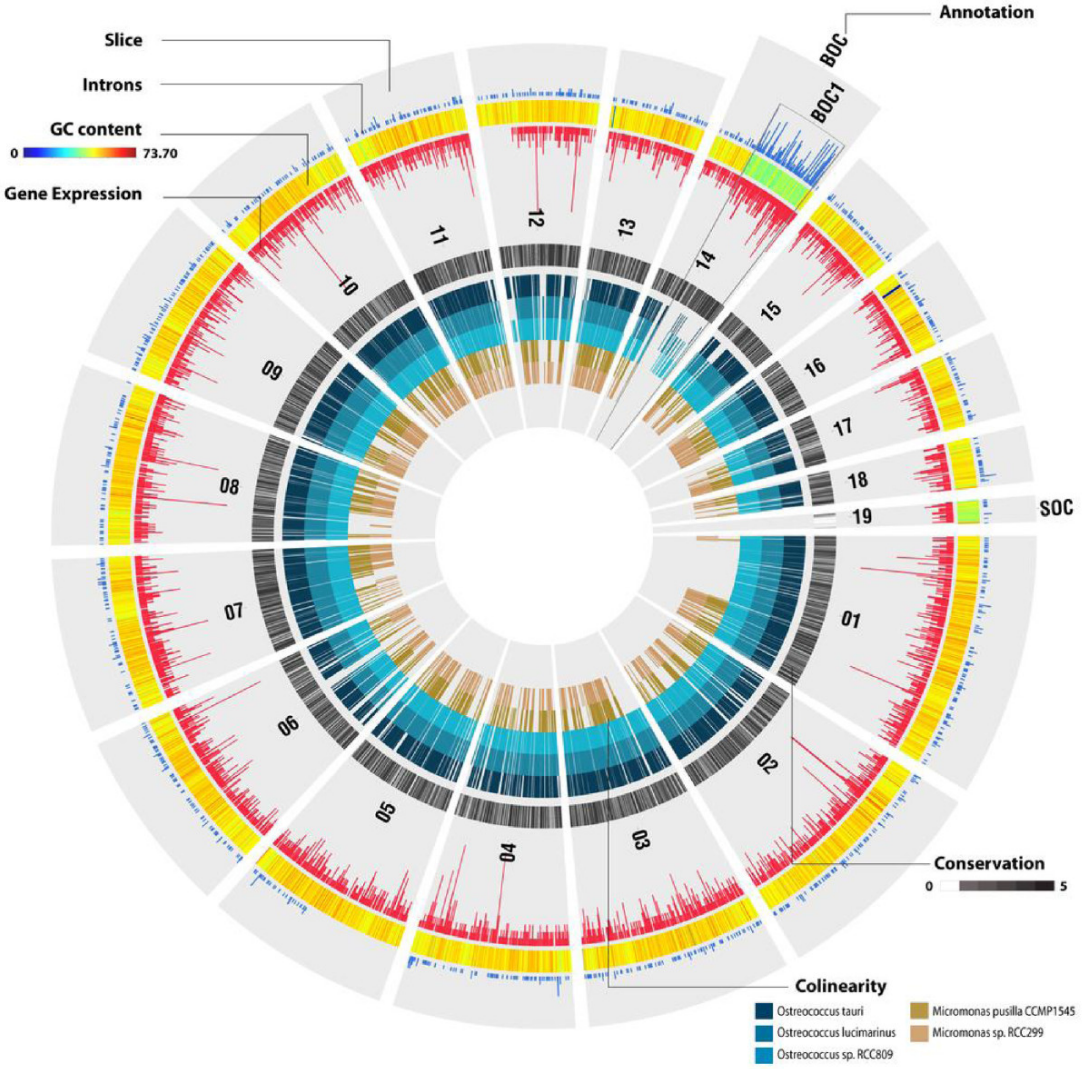


Figure 3.3. Integrative and comparative view of the *Bathycoccus* genome showing both structural (GC content, introns, colinearity) and functional characteristics (gene expression, conservation). 'Slice' represents a single chromosome or region drawn using one gene per unit. 'Introns' and 'Gene Expression' denote the number of introns and uniquely mapped ESTs per gene, respectively. To improve legibility, an upper limit was set for the EST and intron count per gene by removing the top 2%, resulting in a threshold of 12 and 13 for intron count and gene expression, respectively. The GC content is plotted using a window size of 500 bp. The 'Annotation' track represents specific chromosomes or regions denoted by the different grey boxes; BOC and SOC refer to big and small outlier chromosome, respectively. 'Conservation' represents, for each gene, the number of Mamiellales species in which a BLAST hit can be found (E-value threshold 1e-05; range 0 to 5 species). 'Colinearity' shows for each gene if it resides in a genomic region showing colinearity with another Mamiellales species. The circle plot was drawn using the Circos circular visualization software [314].

specific for the BOC1 region, the C-hunter tool *(section 3.4.5)* was applied to detect significant physical clustering of highly expressed genes and intron-containing genes on the different chromosomes. C-hunter analysis revealed that the BOC1 region shows, in all species, a significant over-representation of EST-supported genes. Globally, 75% of all BOC1 *Bathycoccus* genes are EST supported versus 47% for non-BOC1 genes *(Figure 3.3)*. After correcting for the overall 1.6-fold higher expression of BOC1 genes, BOC1 genes related to chromatin assembly, protein transport activity and signal transduction showed increased expression levels. To verify whether the high expression is a property of the low GC genomic BOC1 region (for example, due to a more open chromatin structure [315]), we checked the expression level of the genes on the other low GC chromosome, SOC. We found that SOC genes had no difference in expression level compared to the genes on the 17 other chromosomes. We further investigated whether this higher expression rate is an intrinsic property of the genes

themselves, and estimated the expression levels for orthologs in *Chlamydomonas*, *Volvox* and *Coccomyxa* sp. C-169 *(Supplementary Figure 3.3)*. In all three species, BOC1 orthologs were also more highly expressed than other genes in the genome, suggesting that the higher expression of BOC1 genes in the Mamiellales is related to their function. Alternatively, this pattern might also be due to the global positive correlation, observed for all Mamiellales, between intron content and expression. Although the high expression of basic housekeeping BOC1 gene functions might yield increased metabolic rates and overall growth, it is not clear whether the physical clustering of BOC1 genes in the Mammiellophyceae lineage is based on adaptive gene relocation or constrained ancestral location [316].

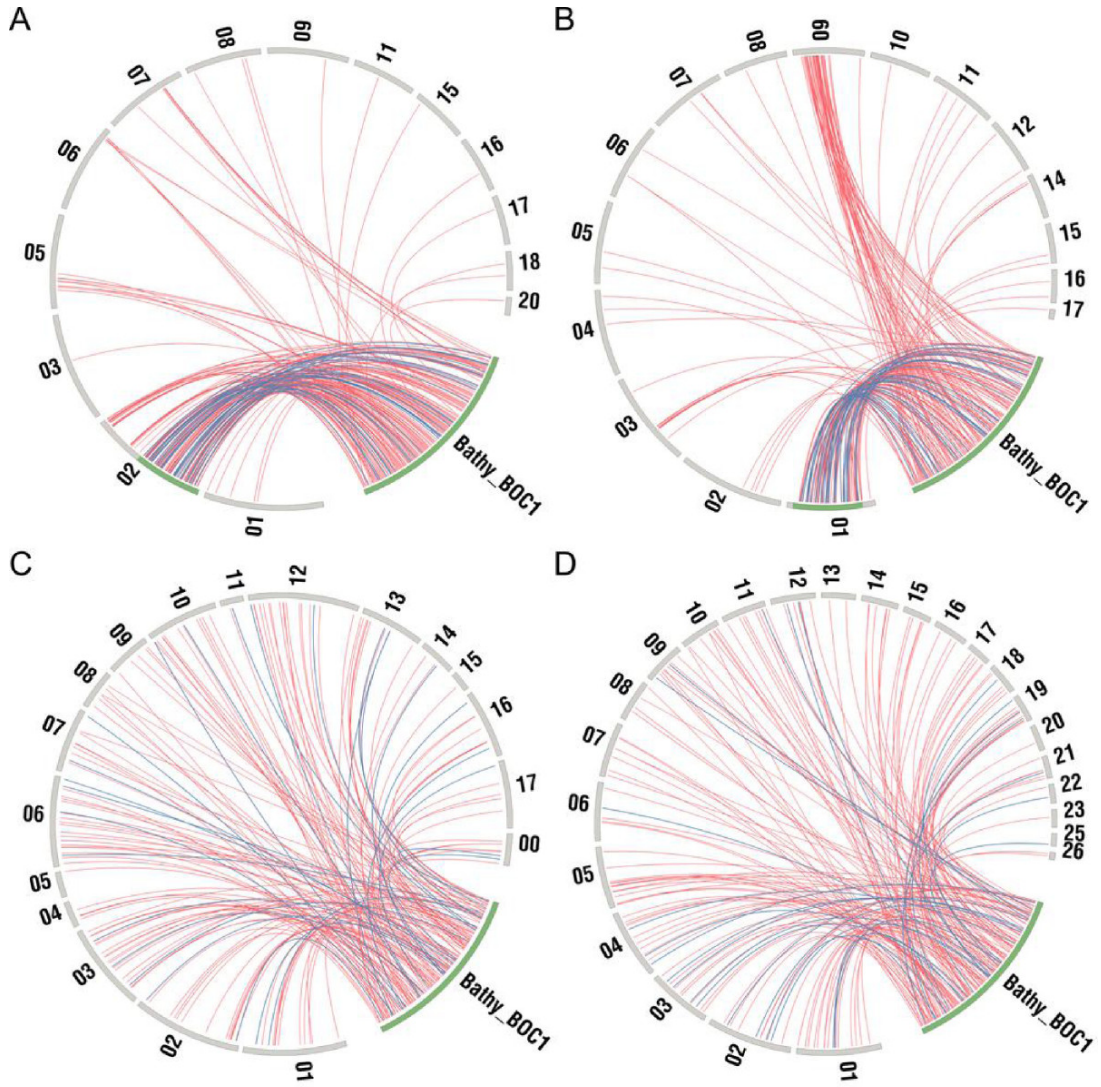


Figure 3.4. Distribution of the *Bathycoccus* BOC1 orthologous genes in the genome of several other green alga species. (a-d). *Bathycoccus* BOC1 orthologous genes in the genomes of *Ostreococcus tauri* (a), *Micromonas* sp. RCC299 (b), *Chlamydomonas reinhardtii* (c) and *Coccomyxa* sp. C-169 (d), each peripheral bar representing a chromosome. The *Bathycoccus* BOC1 region genes (lower right corner, labeled as Bathy_BOC1) are connected by red lines to their orthologs (curated as best BLAST hits) on the chromosomes of other species. Where a *Bathycoccus* gene also represents a BOC1 Mamiellales core gene, the link is coloured blue. Green bars show the BOC1 regions in *Bathycoccus*, *Micromonas* and *Ostreococcus*. The *Micromonas* sp. RCC299 region showing partial clustering of BOC1 orthologs lacks typical BOC1 features. This region has high GC content (66%), is not enriched for a high intron content and does not group highly expressed genes. For the sake of legibility, all small scaffolds of *Chlamydomonas reinhardtii* that harbored a *Bathycoccus* BOC1 orthologous gene were joined together into a virtual chromosome 00 (scaffolds 19, 20, 22, 23, 24, 26 and 32).

The BOC1 region also displays structural specificities that are absent in the rest of the genome. Besides its

low GC content, genes in the BOC1 region are split by many small (40 to 65 bp) AT-rich introns [181, 182]. This feature is present in all of the sequenced Mamiellales genomes *(Supplementary Figure 3.4)* and absent from the genomes of other green algae *(Supplementary Figure 3.4)*. There is no universal RNA-fold for these introns and no conserved sequence motifs (for example, branch points, splice sites) could be detected. Although the only intrinsic indication from their DNA sequences that they are introns comes from their AT-richness relative to the surrounding GC rich exons, their existence is clear from EST data. Consequently, the BOC1 region includes a high proportion of multiple-exon genes, a feature absent in the rest of the genome *(Supplementary Table 3.2)*. In *Bathycoccus*, 103 of the 214 BOC1 genes harbor 330 introns, an intron content tenfold higher than in the rest of the genome (average of 1.54 and 0.15 introns per BOC1 and non-BOC1 gene, respectively).

In conclusion, the BOC1 region in the Mamiellales has unique structural characteristics: it represent one contiguous low(er) GC content region in the chromosome, flanked by two high(er) GC content regions at the extremities and carries between 193 and 633 genes depending on the species examined. The gene order within the region shows little colinearity between species and it encodes a high proportion of often vital housekeeping genes with elevated expression levels clustered together in a pattern unique to the Mamiellales *(Figure 3.4)*. The biological reason for the existence of this region remains obscure, although its structural characteristics (shuffling of genes, small introns, low GC content) concur with the hypothesis that it may be a sex or species-barrier chromosome [317].

3.2.5 The small outlier chromosome in *Bathycoccus*

The size of the *Bathycoccus* Small Outlier Chromosome (SOC) is 146,238 bp. compared to around 150 kb in *Ostreococcus lucimarinus* and 200 to 250 kb in *Micromonas*. The SOC average gene density in *B. prasinos* is slightly lower than that observed in the other chromosomes (72 genes with an average of 2.0 kb per gene in SOC compared to 1.7 kb per gene in the global genome), with a similar expression level based on EST counts. Only 44% of the genes in SOC have a potentially identified function compared to 77% in other chromosomes. Furthermore, up to 75% of the SOC genes have no known plant orthologs, in sharp contrast to most other chromosomes, where most genes share green lineage descent. Last but not least, in *Bathycoccus*, 24 of the 34 SOC genes having an identified function group in two categories. The first group encodes enzymes involved in metabolism of glycoconjugates (17 genes), mainly glycosyltransferase (12 genes), and the second is related to methyl transferases (7 genes). These features are globally similar in the other known Mamiellales SOCs, where the same two dominant gene functions were found *(Table 3.2)*. However, despite their common function, no synteny and almost no orthologous relationships could be established between the SOCs of the different Mamiellales' species, suggesting a more functional convergence than a common phylogenetic origin. To explain the presence of such genes in SOCs, an alien origin of these chromosomes was proposed, which could have yielded some selective advantages in cell surface processes, potentially related, for example, to defense against pathogens or other environmental interactions [182]. However, since SOCs and BOCs have now been found in all sequenced mamiellophycean genomes, it is likely that their lower GC composition, higher proportion of specific genes and higher evolution rates [221, 318] are being maintained by the same evolutionary pressure in all of these species. Interestingly, a paper on the cyanobacteria *Prochlorococcus* describes how variable genomic islands showing similar characteristics to those found in SOCs (low number of orthologs, a high level of horizontal gene transfer (HGT) and a high fraction of sugar-modifying enzymes, methyl transferases and membrane associated proteins) are involved in resistance to viruses [319]. The viral resistance determined by these genomic islands induced a fitness cost measured either by a reduced growth rate and/or a more rapid infection by other viruses. The three genera *Bathycoccus*, *Micromonas* and *Ostreococcus* are the microalgae tested, which are among the most attacked by viruses [320], and viral resistance phenomena showing similar characteristics to what is reported for *Prochlorococcus* (reduced growth rate and higher infection rate by other viruses) have been reported to occur frequently [321]. It is tempting to link this unusual high viral sensitivity and the ability to develop rapid and frequent resistance to these attacks to the presence of SOCs. Interestingly, two other Mamiellales species (*Mamiella* sp. or *Mantoniella squamata*) were tested recently and did not show this high viral sensitivity (N Simon, personal communication). It can be predicted that if our hypothesis on the link between SOC and viral hypersensitivity/resistance is correct, these species should not present a SOC-like structure in their genome.

Species	Chrom. nr	Size (kb)	GC (%)	ORF nr	Gene densities (bp\gene)	Identified genes	Sugar met.	Methylation enzymes	Other function
Bathycoccus sp.	19	146	42	72	2,031	34 (47%)	17 (24%)	7 (10%)	4 (6%)
Ostreococcus lucimarinus	18	149	53	78	1,915	32 (41%)	16 (21%)	5 (6%)	11 (14%)
Micromonas sp. RCC299	17	215	51	80	2,684	30 (38%)	14 (14%)	7 (7%)	9 (11%)

Table 3.2. Characteristics of the small outlier chromosomes for *Bathycoccus* and one *Micromonas* and one *Ostreococcus* species. met, metabolism.

3.2.6 Phylogenomics suggests many horizontal gene transfers

Based on the observation that no plant homologs could be found for many annotated *Bathycoccus* genes, a systematic analysis was performed to unravel their origin. Since plain sequence similarity search strategies are insufficient to reliably trace a gene's evolutionary history [322, 323], a two-step comparative approach was applied to identify putative horizontal gene transfer (HGT) events. After comparing each *Bathycoccus* protein sequence against the National Center for Biotechnology Information (NCBI) protein database, 6,550 phylogenetic trees were constructed and conflicts between the gene and organism phylogeny were determined. Whereas clustering patterns where the nearest neighbour in the tree corresponds with a homolog from a species outside the plant lineage were scored as HGT, in some cases ancestral gene duplication followed by differential gene loss or artefacts of phylogenetic reconstruction methods due to unusual modes of protein evolution could yield misleading results [324]. There were 428 genes (6%) that clustered with a homologous gene from a species outside the green lineage, whereas the remaining genes grouped with Viridiplantae genes (70%) or did not show any significant similarity. Among the 428 putative non-Viridiplantae genes, 80% were of non-green eukaryotic origin while 17% were bacterial orthologs *(Figure 3.5a)*. For the 354 non-green eukaryotic genes, a high proportion came from Metazoa and Stramenopiles (42% and 28%, respectively). Gene Ontology enrichment analysis showed that around 50% of the non-Viridiplantae genes (including prokaryotic genes) were involved in metabolism. Focusing on the most enriched categories revealed genes involved in zinc ion binding (61, 6-fold enrichment), sialyltransferase activity (27, 12-fold enrichment), glycosylation (27, 11-fold enrichment) and ankyrin repeats (5-fold enrichment) (these observations are discussed further in the following section). Application of conservative selection criteria (retaining only phylogenetic trees with bootstrap support >90% and more than 50% protein alignment coverage) yielded 79 genes with non-plant nearest neighbours, which we propose might originate from HGTs, either from eukaryotes (82%) or prokaryotes (18%). Most of these 98 highly probable HGTs (43%) show unknown functions and the others, both originating from pro- or eukaryotes, show metabolite functions (based on similarities with protein domains). The absence of detectable eukaryotic HGT in *Arabidopsis thaliana*, our negative control, suggests that this finding is not an artifact of the method. Using the same approach, previous putative large-scale HGTs have been reported in the available nuclear diatom genomes [325, 326], both from bacteria (784 genes in *Phaeodactylum tricornutum*) or from the green lineage (>1,700 genes). However, although no other 'eukaryotic' potential HGTs are discussed in these papers, orthologous genes shared with other eukaryotic lineages were also described. The presence of green genes in diatoms has been explained by endosymbiotic gene transfers and an alternative hypothesis would be that the presence of stramenopile genes in *Bathycoccus* may reflect an opposite gene flow from diatom-like cells to Mamiellales. This hypothesis seems unlikely, however, because most of the stramenopile genes found in *Bathycoccus* are specific to this species and are not found in other Mamilelalles genomes. Alternatively, this mosaic gene repertoire could be the consequence of (i) parallel or convergent molecular evolution or (ii) the evolution through gene loss of a large ancestral genome, with massive and selective gene losses in all Mamiellales descendants, concurrent with genome reduction. However, this scenario is less parsimonious compared to HGT and, again, seems unlikely because of the phylogenetic breadth of the selectively retained genes (bacterial and from different supergroups of the eukaryotic tree of life).

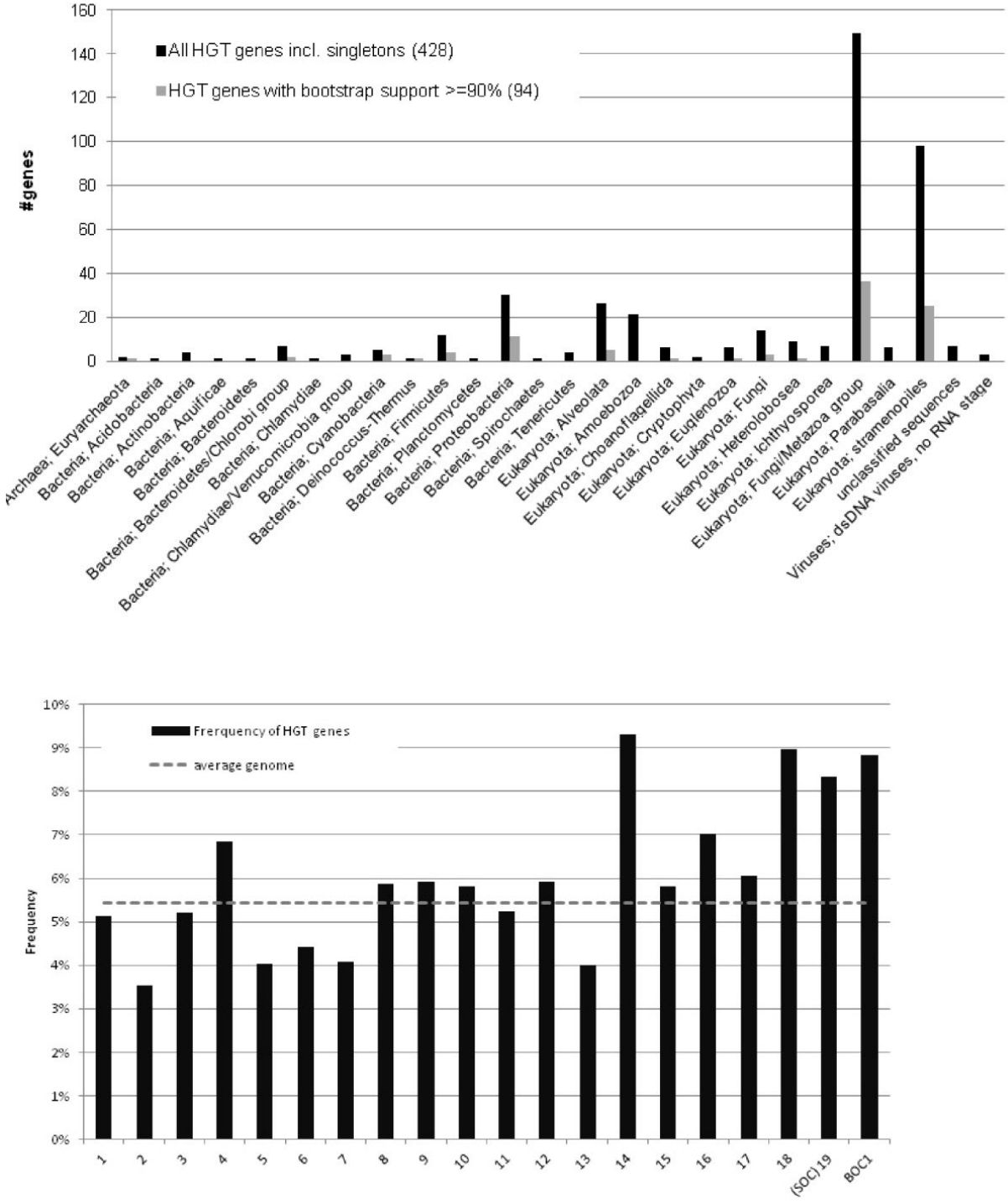


Figure 3.5. Potential horizontal gene transfer in *Bathycoccus*. **(a)** Taxonomic distribution of HGT genes identified using BLAST and by phylogenetic analysis of each gene (excluding genes with a multi-kingdom punctuate distribution). Only taxonomic groups including multiple genes are displayed. **(b)** Frequency of 428 HGT genes over the different chromosomes. The last bin reports the fraction of HGT genes in the BOC1 region (a subset of chromosome 14).

In line with a recent report about the acquisition of ice-binding proteins in sea ice diatoms from prokaryotic origin [327], it is tempting to speculate that the HGT genes contribute new functional properties to the *Bathycoccus* genome. The analysis of a large DNA virus in *Ostreococcus tauri* suggested that the capture of host DNA in viral genomes could represent a mechanism for the transfer of genes between eukaryotic cells [184]. This idea was confirmed by the additional sequencing of four double-strand DNA marine prasinovirus genomes (infecting *Bathycoccus*, *Micromonas*, and *Ostreococcus*), showing that these viruses encode a gene repertoire of certain amino acid biosynthesis pathways never previously observed in viruses that are likely to have been acquired from lateral gene transfer from their host or from bacteria [188]. A similar eukaryotic phytoplankton-virus system was also described in *Emiliania huxleyi*, mediating the transfer of seven genes related to sphingolipid biosynthesis [328].

To verify whether specific genomic regions or chromosomes would be more likely to harbor genes arriving via HGT, we estimated the number of HGT genes per chromosome. We observed that transferred genes were more or less equally distributed over the different chromosomes, except for the low GC outlier chromosomes, which contained higher fractions of HGT genes (BOC1 and SOC contain 1.63 and 1.54 times more HGT genes compared to the genome-wide average) *(Figure 3.5b)*. Different possibilities for the increased abundance of HGT on the outliers include, for example: (1) they may have specific sequence features that can serve to integrate HGT genes that are subsequently re-arranged and embedded in other locations in the genome; (2) it may reflect a lower density of essential gene functionalities in outliers, which could thus support a higher density of random insertions; or (3) there might be a lower level of recombination on these chromosomes, reducing the rate of removal of deleterious alleles via sexual recombination. None of these scenarios are mutually exclusive.

In the *Bathycoccus* genome, the gene copy number is highly expanded for four specific gene families, phenomena not found (or at very low copy number expansion) in other Mamiellales or other algae *(Table 3.3)*. Of these, two are involved in the metabolism of sialic acids, that is, sialyltransferases (69 gene copies) and sialidases (23 gene copies), the two others being ankyrin-repeat proteins (149 gene copies) and zinc finger proteins (48 gene copies) *(Table 3.3)*. Among these 289 gene copies, 105 (36%) are represented within the 428 probable genes acquired by HGT, representing 24% of them.

Gene Family[a]	Copy number					
	Bathycoccus	***Micromonas***		***Ostreococcus***		
	prasinos	**sp. CCMP1545**	**sp. RCC299**	***lucimarinus***	**sp. RCC809**	***tauri***
Glycosyl transferase, family 29 (IPR001675)	78	0	1	0	2	0
HOM000519	43	0	0	0	0	0
HOM002813	10	0	0	0	0	0
HOM005062	10	0	0	0	0	0
HOM007941	6	0	0	0	0	0
Ankyrin repeats (IPR011040)	186	124	107	74	55	67
HOM000035	149	56	9	17	6	6
Sialidase/ neuraminidase (IPR011040)	23	1	0	0	0	0
HOM002557	17	0	0	0	0	0
HOM005056	5	0	0	0	0	0
Zinc Finger, C2H2	53	29	35	19	3	17
HOM000293	48	5	4	1	1	1

Table 3.3. Expanded gene families in the *Bathycoccus* genome. [a]Protein domain description including InterPro identifier. HOM identifiers refer to gene families in pico-PLAZA [329].

3.2.7 Sialic acid metabolism in *Bathycoccus*

The two enzyme families involved in the metabolism of sialic acids are not present in other known green algae genomes, and both gene families are dispersed all along the *Bathycoccus* genome without evident clustering or tandem duplication. Although, on average, 15% of the genes have introns in *Bathycoccus*, no introns (except three genes) were found in any gene from both families. Genes annotated as sialyltransferases correspond to glycosyltransferases family 29 in the CAZy classification, which comprises enzymes able to transfer sialic residues during glycosylation of proteins or lipids [330]. All the *Bathycoccus* sialyltransferases showed a metazoan taxonic affiliation and none of them gave significant hits with bacteria. These enzymes are type II single pass membrane proteins usually known to be anchored in the Golgi membranes [331, 332]. A potential hydrophobic transmembrane domain was detected on the amino-terminal extremities of all the *Bathycoccus* sialyltransferases *(Figure 3.6a)*. For almost all the 69 genes (only 19 are known in human), the sialyltransferase domain is located in the carboxy-terminal part of the protein, whereas the amino-terminal domain is composed of a highly variable stem region *(Figure 3.6a)*. Although the existence of complete and active sialyltransferases in plants is still a matter of debate [332], all four metazoan consensus motifs were found in the *Bathycoccus* genes.

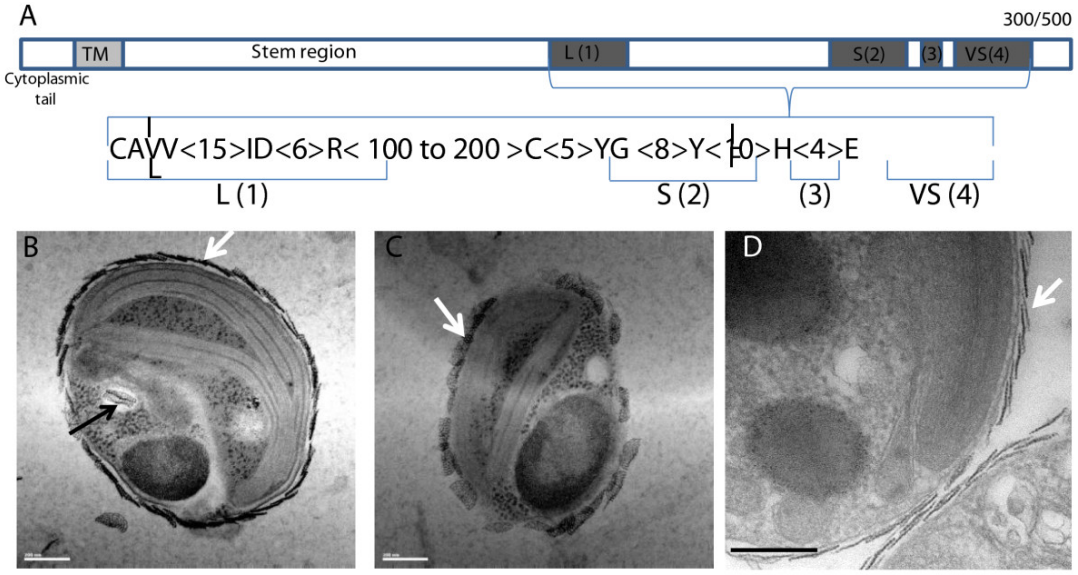


Figure 3.6. Sialyltransferase gene family and external scales covering *Bathycoccus*. (a) Structural organization of the *Bathycoccus* RCC1105 sialyltransferase gene family. TM, transmembrane domain. L(1), S(2), (3) and VS(4) correspond to the four metazoan consensus motifs described for this gene family [331, 332]. Letters in the scheme below are the amino acid one-letter code with alternative possibilities for positions indicated between brackets. **(b-d)** Details of external scales covering *Bathycoccus* RCC1105 cells; white arrows indicate positions of external scales around the plasmic membrane while the black arrow indicates one intracellular scale inside a vesicle.

The second gene family includes sialidases (or neuraminidases), which are enzymes cleaving the terminal sialic acid residues from glycoproteins or glycolipids. Again, this gene expansion is specific to *Bathycoccus*. In contrast to the previous family, no clear domain organization could be defined in sialidases, but some key amino acids known to be involved in the catalytic activity are conserved in the *Bathycoccus* family. The taxonomic origin of the sialidases is less clear than that for the sialyltransferases discussed above, and could correspond to either metazoans or bacteria. For sialidases, scores are globally weak and best blast hits are found mostly with hypothetical proteins either from the choanoflagellate *Monosiga brevicolis* or from the green alga *Chlorella variabilis* (where only one sialidase has been annotated).

The expansion of these two enzyme families prompted us to look for specific potential 'sialic acid' metabolism in *Bathycoccus*. The composition of flagellar scales in *Scherffelia dubia* (phylum Chlorophyta, class Chlorodendrophyceae) was found to be a mix of acidic polysaccharides having similar structures to sialic acids [333].

Although the chemical nature of the scales covering the *Bathycoccus* cell membrane is unknown, it is tempting to establish a correlation between the potential biosynthetic pathway of these scales and the expansion of gene families involved in the metabolism of sialic acids. Furthermore, we confirmed previous electron microscopy studies [165, 334] showing that, in *Bathycoccus* as in other Mamiellophyceae, scale biosynthesis occurs inside intracellular vesicles with striking resemblance to Golgi vesicles *(Figure 3.6b-d)*; that is, in agreement with the notion that they might be produced by sialyltransferases located at the luminal side of intracellular vesicles. Scales almost identical to those of *B. prasinos* are observed in the more closely related *Mantoniella squamata* [335], where they are also extruded to the surface after transport via the Golgi body [336–338].

3.2.8 Other *Bathycoccus* expanded gene families

One of the two other highly expanded gene families in the *Bathycoccus* genome are ankyrin-repeat proteins (149 gene copies). This family is also expanded, although to a lesser extent, in the *Micromonas* strain CCMP1545 (56 copies), whereas only very few copies were detected in other Mamiellales *(Table 3.3)*. These genes have ankyrin repeats located in the carboxy-terminal part of the protein whereas the amino-terminal part has no hit in GenBank. There are also many other ankyrin repeats containing genes in *Bathycoccus* as in both plants and microalgae, but associated with different protein domains that often have predicted functionalities. Indeed, the ankyrin repeat is considered as one of the most common protein-protein interaction motifs in nature [339]. The 149 *Bathycoccus*-specific genes were not distributed randomly among chromosomes, with the bigger chromosomes having few copies, whereas chromosomes 12 or 19 bear many tandem duplicated genes. No obvious function can be attributed to these genes. However, by analogy with the human membrane-associated ankyrin, which is responsible for the attachment of the cytoskeleton to the plasma membrane, it is possible that a number of these genes might function in some way to bind extracellular scales to the plasmic membrane, although experimental evidence is lacking. It has been shown, however, by electron microscopy coupled to immunogold that scales in *Scherffelia dubia* are linked to the membrane by glycoproteins [333]. In addition, in *Tetraselmis striata* (Chlorodendrophyceae) some scale-associated glycoproteins may provide connections between scales and the underlying flagellar membrane [337].

The last group of expanded genes in *Bathycoccus* are zinc finger proteins. There are many zinc finger proteins in microalgae and in plants, but the family specifically expanded in *Bathycoccus* is most related to the C2H2-type zinc finger DNA-binding domain of certain integrases, which share a common alpha/beta two-layer sandwich core structure. The typical organization of the 48 copies identified in the *Bathycoccus* genome *(Table 3.3)* includes a short amino-terminal part (around 20 to 40 amino acids) followed by a strongly acidic region (10 to 20 amino acids) and by 2 to 6 C2H2 domains. Zinc finger proteins were originally identified as DNA-binding domains, although a growing body of evidence suggests an important and widespread role for these domains in protein binding. There are even examples of zinc fingers that support both DNA and protein interactions, and, globally, C2H2 protein-protein interactions are proving to be more abundant than previously appreciated [340].

The most parsimonious explanation for the abundance of the four expanded gene families would be an initial single HGT event followed by expansion in the *Bathycoccus* genome. The potential function of these four gene families and their expansion only in *Bathycoccus* also suggest that they could all be involved in the biosynthesis, exportation and fixation of the scales around the external membrane, and possibly for protection of the cell. Several other members of the Mamiellales have morphologically similar scales around the cells, but they are absent in the two genera *Micromonas* and *Ostreococcus*. The most parsimonious evolutionary scenario to explain these observations is that the scale synthesis pathway was acquired by the ancestor of the Mamiellales (or even before) and has been lost in the two naked genera. This scenario predicts that similar gene family expansions should be found in the genomes of other scaled Mamiellaophyceae but not in *Micromonas* and *Ostreococcus*. This is the case for *Micromonas* and *Ostreococcus*, but the genome sequences of other scaled species are not yet available.

3.3 Conclusions

Mamiellophyceae, and more particularly the three genera *Bathycoccus*, *Micromonas* and *Ostreococcus*, are dominant in different marine areas, where they can play an important role in the primary biomass production.

However, the ecological importance of *Bathycoccus* has probably been overlooked these past years, although it was sporadically mentioned in several studies [212, 304–306]. The availability of this genome, coupled to the development of new sequencing possibilities for metagenomes [307, 308] from various marine environments, opens the door to future comparative studies and to a better understanding of the adaptations of the organisms to their environment.

3.4 Materials and methods

3.4.1 *B. prasinos* RCC1105 genome and EST sequencing and annotation

The sequenced strain *B. prasinos* RCC1105 was isolated in the bay of Banyuls sur mer at the SOLA station. The genomic DNA was extracted from cell pellets containing a collective total of 6.4×10^{10} cells, using a cetyl trimethyl ammonium bromide protocol (adapted from Winnepenninckx *et al.* [341]). The *Bathycoccus* genome was sequenced using Sanger technology on three independent shot-gun libraries with insert sizes of 3 (TK0AAA, vector pcdna2.1 (BstXI)), 10 (TK0AAB, vector pCNS (BstXI)) and 50 kb (TK0ACA, vector pBeloBAC11 (HindIII) and TK0ACB, vector pBeloBAC11 (BamHI)), resulting in 230,496 reads (180 Mb), 118,070 reads (152 Mb) and 10,368 reads (14 Mb), respectively. After trimming, read numbers were 223,577 reads (174 Mb) for the 3 kb library, 112,842 reads (145 Mb) for the 10 kb library and 8,189 reads (11 Mb) for the 50 kb library, and represented a coverage of 22-fold from 330 Mb of sequenced DNA. The data were assembled using the Genoscope pipeline that includes the software Arachne 3.0 [342]. Expressed Sequence Tags (ESTs) were sequenced from a *Bathycoccus* culture grown to log phase (10^7 cells/ml), harvested by centrifugation and the cell pellets were immediately flash frozen in liquid nitrogen. The total RNA was extracted using the TriReagent (Sigma-Aldrich, Saint-Quentin, France) protocol and mRNAs purified using Poly(A)Purist (Ambion-Applied Bioystems, Saint Aubin, France). Complementary DNAs were constructed and cloned using the CloneMiner procedure (InvitroGen, Saint Aubin, France) with some minor modifications. EST sequences were obtained using pyrosequencing technology developed by Roche (Boulogne-Billancourt, France). A total of 253,791 EST reads were processed through the Genoscope EST pipeline. Short (<60 bp) and low complexity sequences were identified and removed. Clustering and assembly of all 251,875 filtered EST reads resulted in 8,370 EST consensus sequences.

The genome was annotated using the EuGene [101, 343] gene finding system with SpliceMachine [102] signal sensor components trained specifically on *Bathycoccus* datasets. The functional annotation resulted from the synthesis of InterPro and the BLASTP hits against the non-redundant UniProt database. Gene Ontology assignments were derived from the InterPro results. GO enrichment analysis was performed using the hypergeometric distribution with Bonferonni correction for multiple hypothesis testing and corrected P-values <0.05 were retained as significant. The resulting database is publicly available at [344] in a format that includes browse and query options and the genome has been submitted to GenBank.

3.4.2 Comparative sequence and expression analysis

Starting from all protein-coding genes from the included species *(Table 3.1)*, only retaining the longest transcript if alternative splicing variants exist, protein sequences were used to construct gene families by applying sequence-based protein clustering. First, an all against all sequence comparison was performed using BLASTP, applying an E-value threshold of 1e-05 and retaining the best 500 hits [345]. Next, the complete sequence similarity graph was processed using Tribe-MCL (mclblastline, default parameters except I = 2 and scheme = 4) to identify gene families. A set of 154 single-copy core gene families was used to construct the phylogenetic tree depicted in *Supplementary Figure 3.1*.

The boundaries of all Mamiellales BOC1 regions were manually delineated based on gene coordinates, gene family information and GC content *(Supplementary Table 3.2)*. For non-Mamiellales, a 'virtual' BOC1 region was created by taking the best BLASTP hit for each *B. prasinos* RCC1105 BOC1 gene. Putative BOC1 Mamiellales core gene families *(Figure 3.4, blue lines)* were identified by first retaining only those families that contain at least one protein for each Mamiellales species. Next, each family was aligned and manually curated. This was done by inspecting and correcting, if necessary, the structural and functional annotation (NCBI BLAST results plus

InterProScan) of all cluster members. For *Ostreococcus* sp. RCC809 no SOC could be identified in the current draft genome assembly *(section 3.5.3)*.

3.4.3 Comparative genomics

To detect co-linearity within and between species, i-ADHoRe 3.0 was used [346] and all chromosomes from all species were compared against each other and significant colinear regions were identified. All gene colinearity can be browsed using the pico-PLAZA comparative genomics platform [329]. i-ADHoRe was run with the following settings: alignment_method gg, gap_size 30, cluster_gap 35, q_value 0.9, prob_cutoff 0.0001, anchor_points 5 and level_2_only false.

EST databases were retrieved from their respective public repositories and mapped on the Mamiellales genomes using GenomeThreader [347] with a minimum alignment score threshold of 0.95 and minimum transcript coverage of 0.89. Only uniquely mapped ESTs were retained and assigned to genes. When an EST with no strand information overlapped with two adjacent genes, it was assigned to the gene with the highest overlap. For the BOC expression analysis global gene, EST counts were first summarized per functional category. In a second stage, expression enrichment was determined by comparing for each functional category the fraction of BOC expressed genes against the overall fraction of BOC expressed genes (denoted 'relative BOC expression enrichment').

3.4.4 Analysis of potential horizontal gene transfer

For each protein-coding gene a BLAST sequence similarity search was performed against the NCBI protein database, which contains the proteins of all sequenced *Ostreococcus*, *Micromonas* and *Chlamydomonas* species (E-value <1e-05). Starting from a selection of BLAST hits a phylogenetic approach was used to identify the putative origin of all genes. Briefly, good hits (20% top hits relative to the best Bit score excluding query self-hits) were retained per gene, protein sequences and detailed taxonomic information was retrieved, a multiple sequence alignment was generated using MUSCLE and a maximum likelihood phylogenetic tree was constructed using PhyML (100 bootstrap sets, WAG model, kappa estimated, 4 substitution rate categories, gamma distribution parameter estimated, BIONJ starting tree, no topology, branch lengths and rate parameter optimization). For each query gene the corresponding tree topology was investigated to identify the nearest neighbour gene/clade, including bootstrap support, and determine the nearest neighbour taxonomic information. Genes showing complex punctuate patterns [348] (that is, clustering with homologs from different phyla outside the Viridiplantae) were excluded. Singletons refer to genes for which no phylogenetic analysis could be done because they only have a single BLAST hit based on the 20% top hits. Nearest neighbours with bootstrap support >90% and gene coverage of 50% or more in the multiple sequence alignment (MSA) were scored as reliable HGT genes to estimate the fraction of eukaryotic origin. Although the low number of HGT genes found in *Arabidopsis* does not serve as a perfect negative control for the detection of HGT in unicellular green algae, it suggests that, when applied to a full set of proteins of a specific organism, this approach gives a conservative estimate of putative transfer events with a low number of false positives. To verify if, for some HGT genes, homologous genes exist in other algae that were missed during the process of gene annotation, a systematic sequence similarity search (using tblastn, E-value threshold 1e-05 against intergenic sequences of *O. tauri*, *O. lucimarinus*, *Ostreococcus* RCC809, *M. pusilla* and *C. reinhardtii*) revealed that, on average, no homologous locus could be found for 93% of the HGT genes.

3.4.5 C-hunter analysis

Four functional categories (two types with two subdivisions each) were defined and genes were assigned to each class, if applicable. The first type of functional category describes the expression state of a gene (based on uniquely mapped ESTs; is a gene expressed (number of ESTs >0) or highly expressed (number of ESTs >2)) while the second type describes the intron content of a gene (contains an intron (number of introns >0) or contains a 'lot' of introns (number of introns >2)). C-hunter [349] software was used to identify, in all genomes, significant clusters of genes belonging to one of the four functional categories. The C-hunter thresholds for each category

subdivision were determined by reviewing the average expression and intron content of all Mamiellales genes. C-hunter was run with the following parameters: <C-hunter categories.go genome.index genome.go 2 80 80 0.001 50 T chunter output>.

3.5 Supplementary Information

This section contains selected segments of supplementary methods, figures and tables most relevant to this chapter and the topic of this dissertation.

3.5.1 Genome annotation and transposable elements detection

The data sources used to complement the *ab initio* part of EuGene were composed of *B. prasinos* RCC1105 Expressed Sequence Tags (ESTs), protein databases (TAIR10, *O. lucimarinus* proteome and SwissProt), and the other Mamiellales raw genomic sequences. Repeats were detected using RepeatMasker [73] (low-complexity regions and simple repeats + the RepBase library [350]), findpat [351] (exact repeats>40nt), LTRharvest [352] +LTRdigest [353], LTR_seq [354], a BLASTP against all TE-related NRPROT proteins (E-value threshold 1e-05) and a detailed HMMer scan using all profiles from the Gypsy Database [355]. Noncoding genes were detected using an ensemble approach of RepeatMasker [73], RNAmmer [114], tRNAscan-SE [113], INFERNAL [356] and BLASTN (using *O. tauri* RNA data).

3.5.2 Phylogenetic position *Bathycoccus prasinos* RCC1105

Based on phylogenetic profiles present in the pico-PLAZA database [329], which represent the number of gene copies per family and per species, 154 families that were single-copy in 10 sequenced green algal genomes and the outgroup species *Arabidopsis thaliana*, *Oryza sativa* and *Physcomitrella patens*, were extracted. For every single-copy core gene family, a multiple sequence alignment was created using MUSCLE [357]. Alignment columns containing gaps were removed when a gap was present in ≥10% of the sequences. Alignment columns containing gaps were removed when a gap was present in ≥10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the MSA was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. Next, the median value for all these values was calculated. If this median was ≥0, the column was considered as containing homologous amino acids. The different edited multiple sequence alignments were concatenated into one super-alignment using a custom Perl script (35,431 amino acids) and used to construct a phylogenetic tree *(Supplementary Figure 3.1)* using PhyML (100 bootstrap sets, WAG model, kappa estimated, 4 substitution rate categories, gamma distribution parameter estimated, BIONJ starting tree, no topology, branch lengths and rate parameter optimization) [358].

3.5.3 Analysis of SOC in *Ostreococcus* sp. RCC809

From the current RCC809 genome assembly, the most likely SOC scaffold would be chromosome_18. However, it contains a large colinear region with chromosome 10 of *Ostreococcus tauri*, a feature that does not fit with the description of SOCs in the other *Ostreococcus* genomes. The definitive nature of the RCC809 SOC therefore remains speculative.

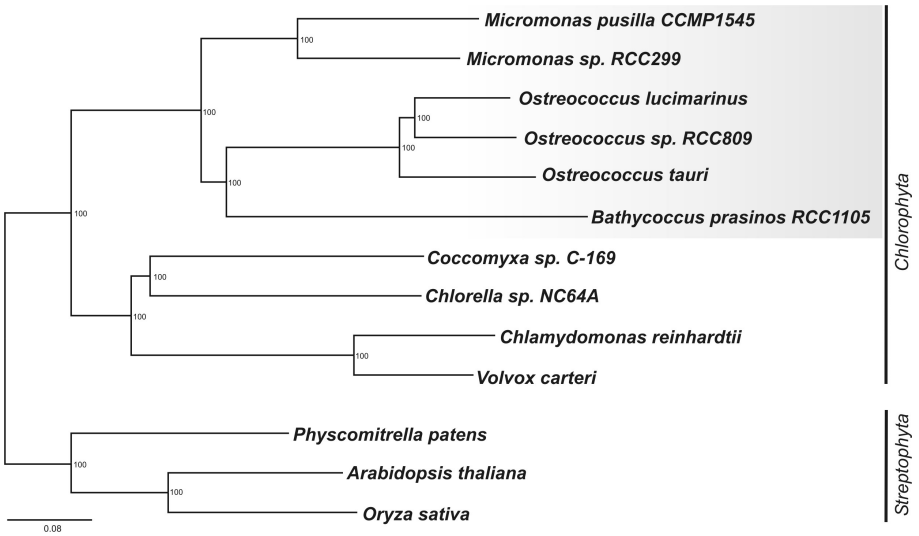
3.5.4 Supplementary Figures & Tables

Genome	Information				
	22 contigs for 19 chromosomes, 1 chloroplast and 1 mitochondrion				
	Genome length: 15,122,588 nt				
	N50*: 8				
	L50*: 937,610 nt				
	Gaps (N>20): 22				
	Total gap length: 36,954 nt				
Genes	**Gene Type**	**Total genes**	**Nuclear**	**Mitochondrion**	**Chloroplast**
	Coding	7,919	7,826	41	52
	tRNA	57	17	26	14
	rRNA	10	4	4	2
	Total	7,986	7,847	71	68
	Gene property	**Number of Genes (% of total Genes)**			
	Multi-exon	1174 (14.70%)			
	EST-support	3692 (46.33%)			
	Homology-support	6789 (85.01%)			
	InterPro domains	6160 (77.13%)			
	GO-labels	3597 (45.04%)			

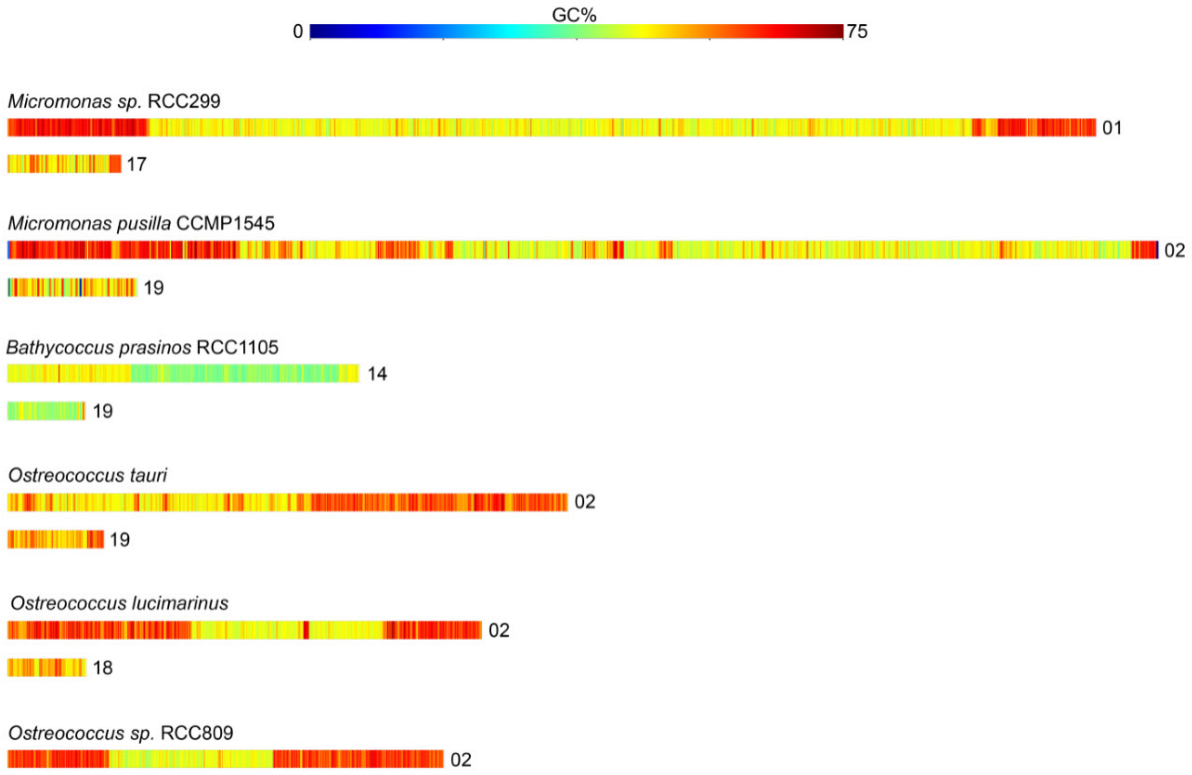
Supplementary Table 3.1. General annotation statistics for *Bathycoccus prasinos* RCC1105. *L50, length of the scaffold that separates the top half (N50) of the assembled genome from the remainder of the smaller scaffolds, if the sequences are ordered by size. N50 is the number of scaffolds that represent the top half of the assembled genomes, if the sequences are ordered by size.

Species	chromosome	BOC1 start	BOC1 end	Length (bp)	GC%
Bathycoccus sp. *prasinos*	14	236365	624661	388296	39
Micromonas sp. RCC299	1	263000	1817000	1554001	47
Micromonas sp. CCMP1545	2	438300	2112000	1673701	48
Ostreococcus lucimarinus	2	345000	709200	364201	47
Ostreococcus sp. RCC809	2	180000	500000	320001	46
Ostreococcus tauri	2	1	575000	575000	50

Supplementary Table 3.2. Annotation of the BOC1 region in different Mamiellales species.



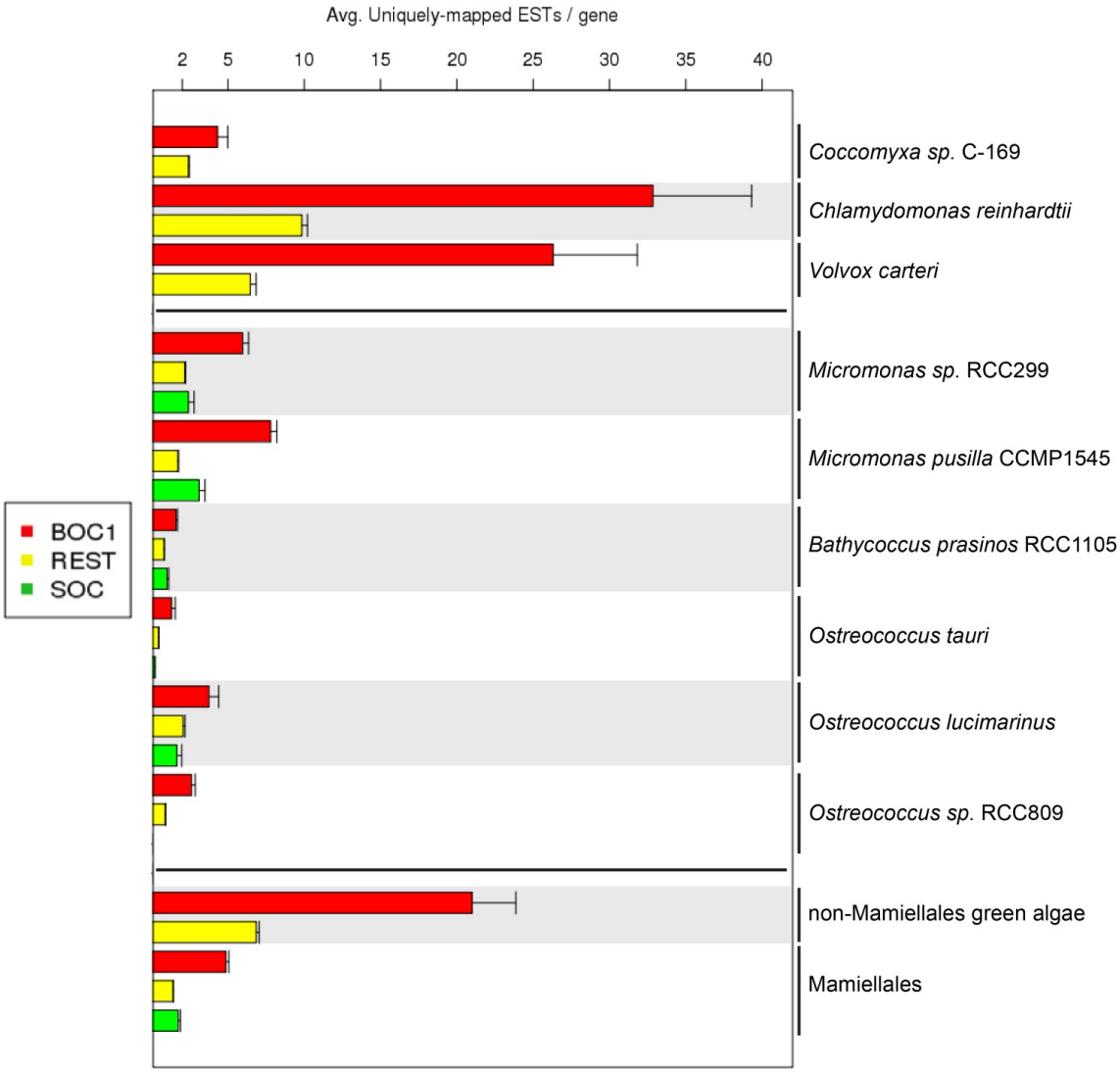
Supplementary Figure 3.1. Maximum likelihood tree depicting the phylogenetic position of *Bathycoccus* RCC1105. A total of 154 single-copy genes conserved in 13 species including plants were concatenated and aligned over 35,431 amino acid positions to construct the phylogeny tree using MUSCLE and PhyML (see details in *section 3.5.2*). Species in the order Mamiellales are indicated by the grey box.



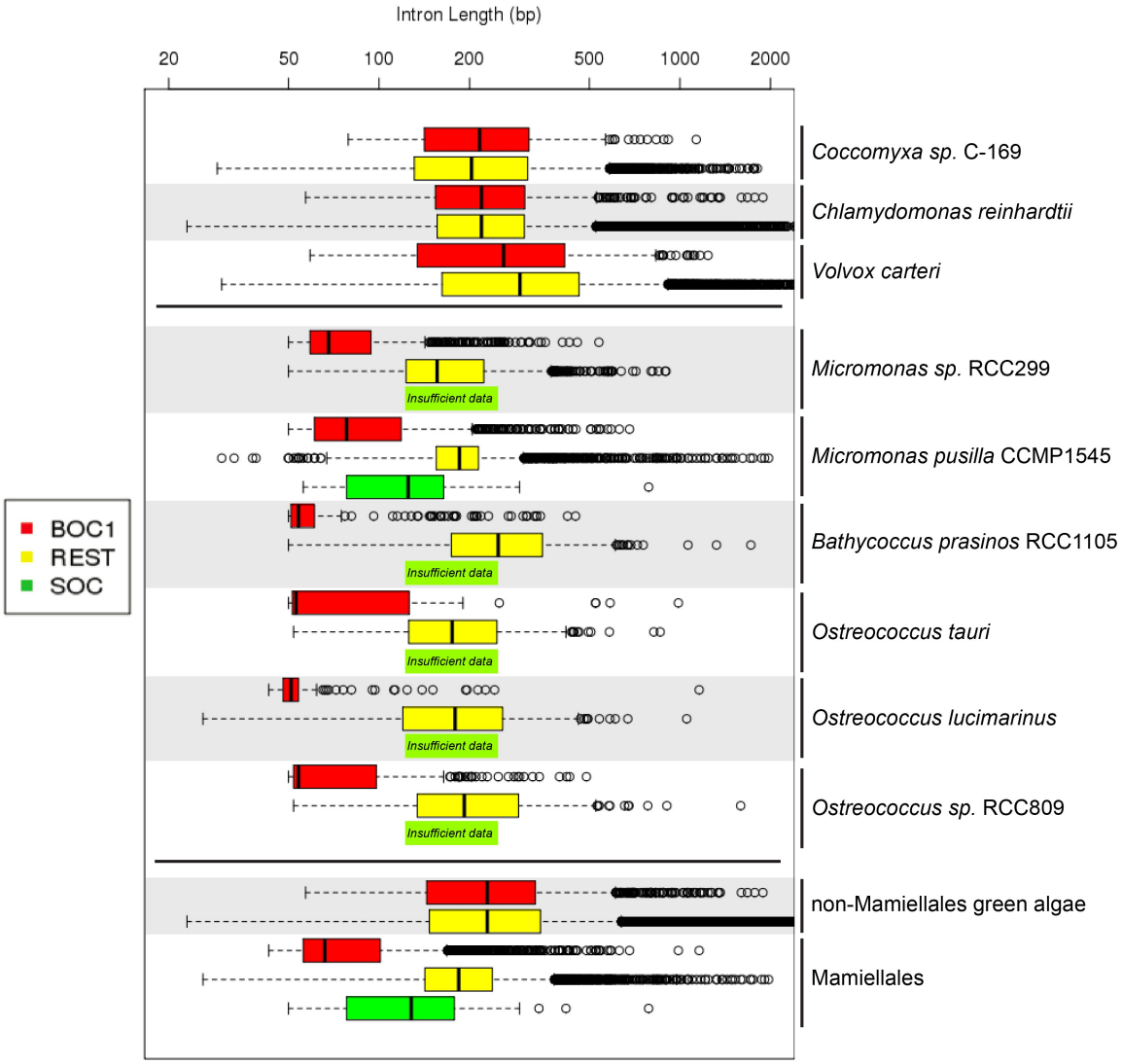
Supplementary Figure 3.2. GC content of outlier chromosomes in Mamiellales genomes. The GC content is plotted using a window size of 2kb. The numbers at the end of each bar indicate the chromosome number. We define the BOC1 region in *Bathycoccus* as that spanning nucleotide positions 236,365 to 624,661.

Locus_id	Functional description
Bathy14g01300	beta-adaptin-like protein C
Bathy14g01380	TFIID component TAF4
Bathy14g01390	Phosphotyrosyl phosphatase activator, PTPA
Bathy14g01470	U3 small nucleolar RNA-associated protein 18
Bathy14g01520	arginyl-tRNA synthetase
Bathy14g01530	glycosyltransferase family 28 protein, putative Monogalactosyldiacylglycerol synthase
Bathy14g01650	Mg-protoporyphyrin IX chelatase
Bathy14g01670	Phosphatidic acid Phosphatase-related protein
Bathy14g01700	glycosyltransferase family 4 protein, putative alpha-1,3-mannosyltransferase ALG2
Bathy14g01860	Caf1 CCR4-associated (transcription) factor 1
Bathy14g02130	ribosome biogenesis protein RLP24
Bathy14g02140	coatomer protein gamma-subunit
Bathy14g02190	CycK-related cyclin family protein
Bathy14g02270	eukaryotic translation initiation factor 4E
Bathy14g02340	histidinol-phosphate aminotransferase, chloroplast precursor
Bathy14g02350	transcription factor IIa large subunit 3
Bathy14g02360	MAK16-like protein
Bathy14g02380	Isoleucine-tRNA synthetase, probable
Bathy14g02640	ATP synthase beta chain, mitochondrial precursor
Bathy14g02730	V-type proton ATPase subunit d 1
Bathy14g02790	60S ribosomal protein L36
Bathy14g02810	U3 small nucleolar RNA-associated protein 6
Bathy14g03000	Ribosome biogenesis protein BOP1
Bathy14g03050	UphC Sugar phosphate permease, putative regulatory protein
Bathy14g03060	1-deoxy-D-xylulose-5-phosphate (DXP) synthase, plastid precursor
Bathy14g03100	Tim circadian rhythm control protein Timeless homolog
Bathy14g03180	Conserved oligomeric Golgi complex component 4
Bathy14g03200	eukaryotic translation initiation factor 6
Bathy14g03330	RNA Polymerase subunit 2

Supplementary Table 3.3. *Bathycoccus* BOC1 Mamiellales core genes and their functional description.



Supplementary Figure 3.3. Gene expression of BOC1, Rest and SOC genes in Mamiellales and non-Mamiellales green algae. For non-Mamiellales, a virtual BOC1 region was created by grouping all the best BLASTP hits for each *Bathycoccus prasinos* RCC1105 BOC1 gene. REST refers to genes not belonging to BOC1 and SOC, respectively. This procedure could not be repeated for SOC, as this region contains too many species-specific genes. Error bars indicate SE.



Supplementary Figure 3.4. Intron length distribution in Mamiellales and non-Mamiellales green algae. For each organism, the lengths of BOC1, REST and SOC EST-confirmed introns are shown. For the BOC1 definition and SOC absence in non-Mamiellales, see *Supplementary Table 3.2 and Supplementary Figure 3.2*. 'Insufficient data' indicates either an absence of EST-confirmed introns or too few data points (less than 11) to construct a boxplot. The data clearly shows that SOC genes carry little (EST-confirmed) introns. For the sake of visibility, intron length outliers above 2000bp are not displayed.

CHAPTER 4

THE COMPLEX INTRON LANDSCAPE AND MASSIVE INTRON INVASION IN A PICOEUKARYOTE PROVIDES INSIGHTS INTO INTRON EVOLUTION

***Bram Verhelst**, Yves Van de Peer and Pierre Rouzé*

Abstract

Genes in pieces and spliceosomal introns are a landmark of eukaryotes, with intron invasion usually assumed to have happened early on in evolution. Here, we analyse the intron landscape of *Micromonas*, a unicellular green alga in the Mamiellophyceae lineage, demonstrating the coexistence of several classes of introns and the occurrence of recent massive intron invasion. This study focuses on two strains, CCMP1545 and RCC299, and their related individuals from ocean samplings, showing that they not only harbor different classes of introns depending on their location in the genome, as for other Mamiellophyceae, but also uniquely carry several classes of repeat introns. These introns, dubbed Introner Elements (IEs), are found at novel positions in genes and have conserved sequences, contrary to canonical introns. This IE invasion has a huge impact on the genome, doubling the number of introns in the CCMP1545 strain. We hypothesize that each IE class originated from a single ancestral IE that has been colonizing the genome after strain divergence by inserting copies of itself into genes by intron transposition, likely involving reverse splicing. Along with similar cases recently observed in other organisms, our observations in *Micromonas* strains shed a new light on the evolution of introns, suggesting that intron gain is more widespread than previously thought.

Contributions

- Intron classification, localisation and landscape
- Detection of Presence/Absence Polymorphisms & metagenomic analysis
- Set-up and maintenance of the *Micromonas*-sections (CCMP1545 and RCC299) on the ORCAE platform
- Figures *(Figures 4.1 to 4.5 and 4.7 and Supplementary Figures 4.1 to 4.8)*
- Tables *(Table 4.1)*
- Writing the manuscript

4.1 Introduction

Recently, several whole-genome sequences have been reported for Mamiellophyceae, eukaryotic picoalgae at the basis of the green lineage that play a major trophic role in the marine environment. Among these are the genome sequences of two *Micromonas* strains, isolated from tropical (Equatorial Pacific; strain RCC299) and coastal waters (Plymouth, English Channel; strain CCMP1545) [221]. One striking outcome of the genome analysis of these algae was the observation of a complex intron landscape in *Micromonas*, especially in the CCMP1545 strain *(Figure 4.1a)*. In common with other Mamiellophyceae, both *Micromonas* strains RCC299 and CCMP1545 feature two distinct classes of introns, corresponding to the unique genome heterogeneity of these picoalgae [1]. At most chromosomal locations, mamiellophycean genes harbor no or few canonical spliceosomal introns with conserved splice sites and branch-point motif [1, 173, 182]. However, in all mamiellophycean genomes studied so far, two low-GC% regions can be identified that harbor peculiar introns [1, 173, 182, 221]. One of the low-GC% regions is located on a chromosome denoted as Big Outlier Chromosome (BOC) and is represented by chromosome 2 in CCMP1545 and chromosome 1 in RCC299. This BOC displays intron heterogeneity with numerous small AT-rich introns in the low-GC% region, dubbed BOC1 introns [1, 359] *(Figure 4.1b)*. A small portion of these BOC1 introns feature noncanonical splice sites. Additionally, in *Micromonas* CCMP1545, Worden *et al.* [221] reported the occurrence of repeat introns, dubbed Introner Elements (IEs). These IEs could be further subdivided into four different families (IE-A1 - IE-A4) based on the presence or absence of specific IE sequence motifs and seemed to be absent from RCC299 or any other published mamiellophycean genome.

In this study, we present an in-depth analysis of these Introner Elements and the discovery of three additional classes: IE-B and IE-D in CCMP1545 and IE-C in RCC299. All four classes show a high degree of within-class sequence conservation, are found on the sense strand of genes, follow similar genomic distribution patterns, and are found at unique positions in genes. These observations stand in sharp contrast to canonical spliceosomal introns, which generally display a very low degree of sequence conservation and are often found at conserved positions in genes. Based on the structural characteristics of IEs and the distribution of their occurrence, we propose that the mechanism by which they replicate possibly involves reverse splicing at the pre-mRNA level and conclude that the replication of IEs provides an important mechanism of intron gain. As a consequence, intron gain could be more widespread than commonly believed.

4.2 Results and discussion

4.2.1 Intron classification

Micromonas introns can be classified into two categories, namely singleton introns, which are all unique in the sense that they do not show significant similarity to other introns in the genome, and IEs, which are a copy of or at least show partial similarity to several or many other introns. To the first category *(Table 4.1)* belong classes that are present in all Mamiellophyceae: the canonical introns and the BOC1 introns. The canonical spliceosomal introns of strains RCC299 and CCMP1545 favor the donor consensus sequence *AG|GTGCGT (Supplementary Figure 4.1)* and have a predicted *NCTGAC* branch-point motif at 43-52 bp upstream of the acceptor site. Comparative intron analysis revealed that 47% of all canonical intron positions are shared between CCMP1545 and RCC299 orthologs, a number that illustrates the divergence of these strains, which are members of different clades (RCC299: clade-II; CCMP1545: clade-V [221, fig. 3]), and probably should be regarded as separate species.

BOC1 introns share few common features, such as their short length and low GC% *(Figure 4.1)*. The majority of BOC1 introns follow the common *GT-AG* splice site rule but have no discernible branch-point motif. Presumably, the drop in GC% *(Figure 4.1b)* across the splice site aids recognition by the splicing machinery. Furthermore, 34 of these introns feature noncanonical *TG* or *CG* acceptor sites, of which the majority is validated by EST alignments. Similar noncanonical acceptor sites have been found in non-prasinophytes as well [361]. Most of the BOC1 intron positions (73%) are shared between both isolates. This percentage is considerably higher than the one for canonical introns (47%), which might be related to a constraint on the BOC1 genes, which are more highly expressed and more often functionally conserved [1].

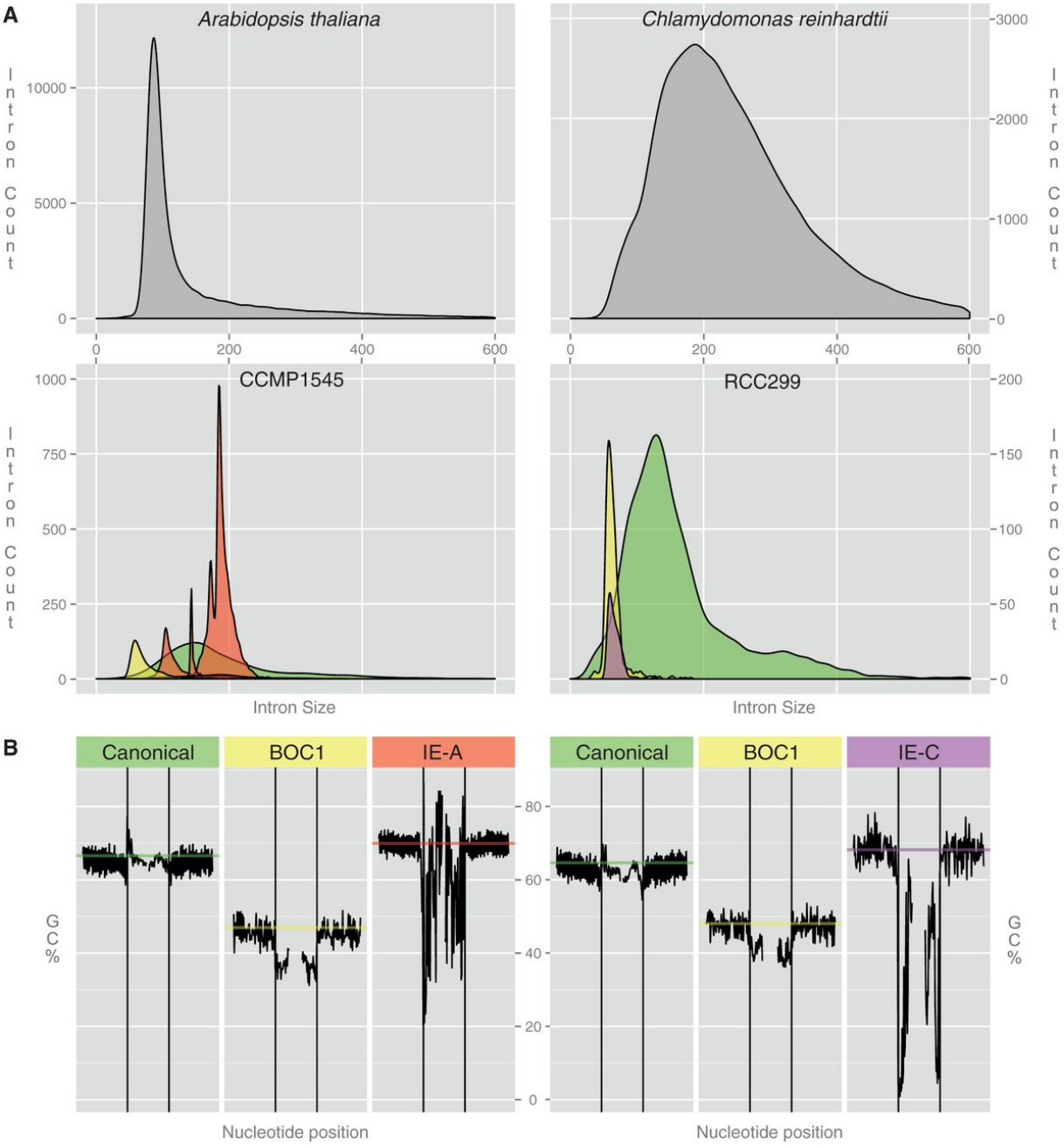


Figure 4.1. The intron landscape of *Micromonas*. (a) Size distribution of different intron classes in *Micromonas* strains CCMP1545 and RCC299 (intron sizes binned per 5 nt). The two panels at the top represent reference intron distributions for *Arabidopsis thaliana* and *Chlamydomonas reinhardtii*. The different classes are canonical (green), BOC1 (yellow), IE-A (red), and IE-C (purple). Due to their low occurrence, members of classes IE-B and IE-D are not displayed. Introns longer than 600 nt are excluded. **(b)** Average GC% of *Micromonas* introns (left: CCMP1545; right: RCC299) and their bordering exon regions. Exon/intron boundaries are marked by black vertical lines, while horizontal lines represent the average GC% of all coding sequences containing at least one intron of the specified class. Exons and introns were trimmed by 3 and 6 nt, respectively, on either end to omit splice-site signals. Only 80 (exon) and 40 (intron) nt on either side of the exon/intron boundary are displayed. Plots were drawn using ggplot2 [360].

Organism	Intron Type	Intron Class (Family)	Intron Count	Avg. Length (nt)	% EST validated	Hosting Genes
RCC299	Singleton	Canonical	4,063	162	31.8	3,063
	Singleton	BOC1	625	65	82.7	157
	Repeat	IE-C	221	67	23.1	150
CCMP1545	Singleton	Canonical	3,553	192	42.6	2,742
	Singleton	BOC1	770	74	90.6	138
	Repeat	IE-A	6,112	173	23.4	3,162
	Repeat	IE-A1	4,328	189	25.0	2,677
	Repeat	IE-A2	1,004	110	16.7	610
	Repeat	IE-A3	328	148	24.4	297
	Repeat	IE-A4	100	185	26.0	93
	Repeat	IE-A?	352	183	21.1	311
	Repeat	IE-B	25	1,830	20.0	25
	Repeat	IE-D	6	374	0.0	6

Table 4.1. *Micromonas* Intron Properties.

4.2.2 Introner Elements

After careful analysis and reannotation of the *Micromonas* genomes, we have identified four distinct classes of IEs *(Table 4.1)*, that is, introns that are repeat elements in strains RCC299 and CCMP1545. These four IE classes differ in terms of host, abundance, sequence, and length. CCMP1545 contains three IE classes: IE-A, IE-B, and IE-D. IE-A has 6,112 members that can be further divided into four families of different size (IE-A1: 4,328; IE-A2: 1,004; IE-A3: 328; IE-A4: 100) and 352 elements with unclear class assignment due to the presence of insertions or deletions (indels) and sequence degeneracy. IE-A sequences consist of a series of sequence motifs, some of which are universal to all IE-A sequences and some of which are specific to one of the subclasses of IE-A *(Supplementary Figures 4.2 to 4.5)*. IE-A members also have very typical splice donor sites, *AG|GYGCGT* or *AG|GTGAGAC*, with the first occurring in IE-A1 and IE-A2, while the latter is almost exclusively found in IE-A3 and IE-A4 sequences *(Supplementary Figures 4.1 to 4.5)*. Fifty-three percent of IE-A1 sequences contain a GC splice donor, a characteristic that was noted in earlier studies but was never linked to the presence of IEs [362]. Overall, IE-A dominates the intron landscape as it represents over half of all introns and is the main cause for the 1 Mb surplus in CCMP1545 genome size over RCC299.

Besides the IE-A introns, there are 463 IE-A-like repeats, which are positioned outside introns or inside pre-existing introns (discussed later). These are remnants of IE-A introns: highly degenerated, partial copies that most often only consist of a small 50-nt motif (motif-C, *(Supplementary Figures 4.2 to 4.5)*, having lost both splice sites and all other motifs crucial for the splicing process. They are found in close proximity to coding sequences (~UTR regions) or within canonical intron sequences, but never in coding sequences where they are counter-selected for to maintain gene functionality.

The IE-B and IE-D class consist of 25 and 6 members, respectively, which have a very variable length, ranging from 100 up to 6,494 nt for certain IE-B members *(Figure 4.2)*. Their *GT-TG* splice sites are highly unusual but have been reported before in other species, including human [363]. Eight of the IE-B sequences harbor a long >3,000-nt open reading frame on the complementary strand. This intron-encoded protein (IEP) lacks homology to other known proteins, except for a small OTU-like protease domain. As such, the function of this protein, or the reason why it is embedded within these IEs, is unknown. The IE-B class contains both the longest documented mamiellophycean intron and the first documented occurrence of a nuclear intron-encoded protein within Mamiellophyceae. A defining characteristic of these two classes is the preference for phase-2 (i.e., the intron sits in between the second and third base of a codon), which contradicts the theory that newly gained introns prefer phase-0 (i.e., the intron sits in between two codons) *(Supplementary Figure 4.6)* [255]. Although

sharing common splice features, IE-Bs and IE-Ds do not show any sequence similarity, which is why they have been ascribed to different classes.

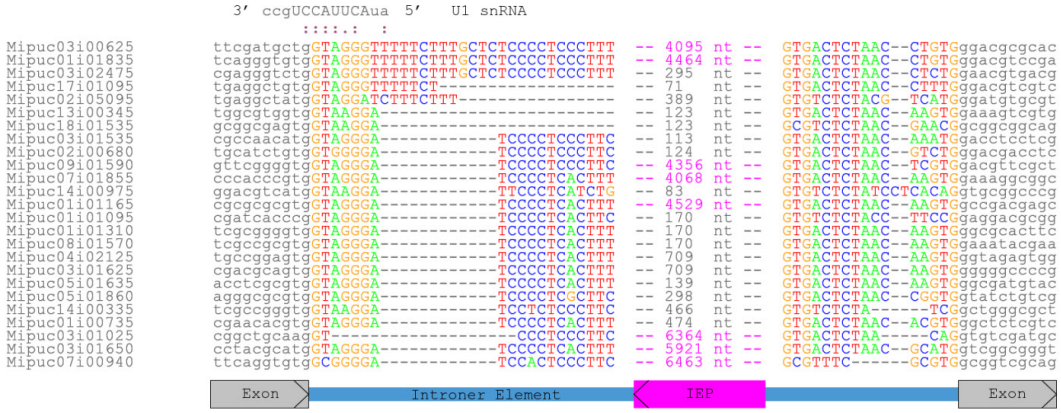


Figure 4.2. Alignment of all 25 IE-B sequences showing the splice site regions in detail. The structure and orientation of exonic regions (grey) and intron-encoded proteins (purple) is represented schematically beneath the alignment. Base-pairing information regarding the donor site (U1 snRNA) is also provided.

As stated previously, the IE-C class (221 occurrences) exists exclusively in RCC299. The IE-C sequences (with an average length of 67 nt) are much shorter than the IEs found in CCMP1545 and feature a highly conserved branch-point motif – *GACTGACG* – identical to the extended branch-point sequence reported for canonical *Ostreococcus* introns [359] *(Figure 4.3)*.

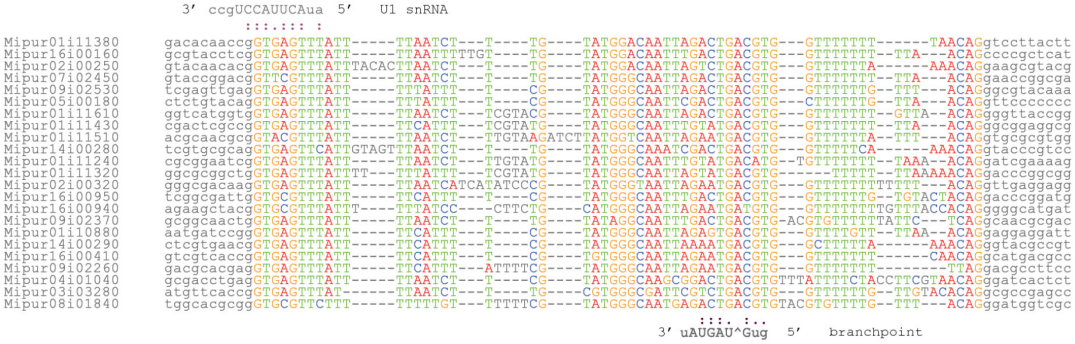


Figure 4.3. Alignment of typical IE-C sequences. Alignment of typical IE-C sequences, flanked by their exonic regions (grey). Base-pairing information regarding the donor site (U1 snRNA) and the branch-point (U2 snRNA) is also provided.

IEs, present in a third of all CCMP1545 genes, are fully functional spliceosomal introns. Beside the fact that they feature the necessary splicing-related motifs (donor and acceptor sites, branch-point, poly-Y tract), EST evidence confirms their excision from primary transcripts. Even more, the non-excision of IEs from the transcripts would generally lead to a premature stop codon resulting in nonsense-mediated mRNA decay [364].

4.2.3 Genomic localization

IEs are not evenly distributed in the genome and are virtually absent from low-GC% areas, such as the AT-rich fraction of the BOC *(Figure 4.4a)*. A second, so-called Small Outlier Chromosome (SOC), low in GC% and found in all mamiellophycean species reported so far, is also completely devoid of IEs. Other chromosomes tend to have the IEs distributed over their entire length, however with reduced densities in regions with lower GC% *(Figure 4.4b)*. Their tendency toward high-GC% areas even surpasses canonical introns *(Figure 4.1b)*.

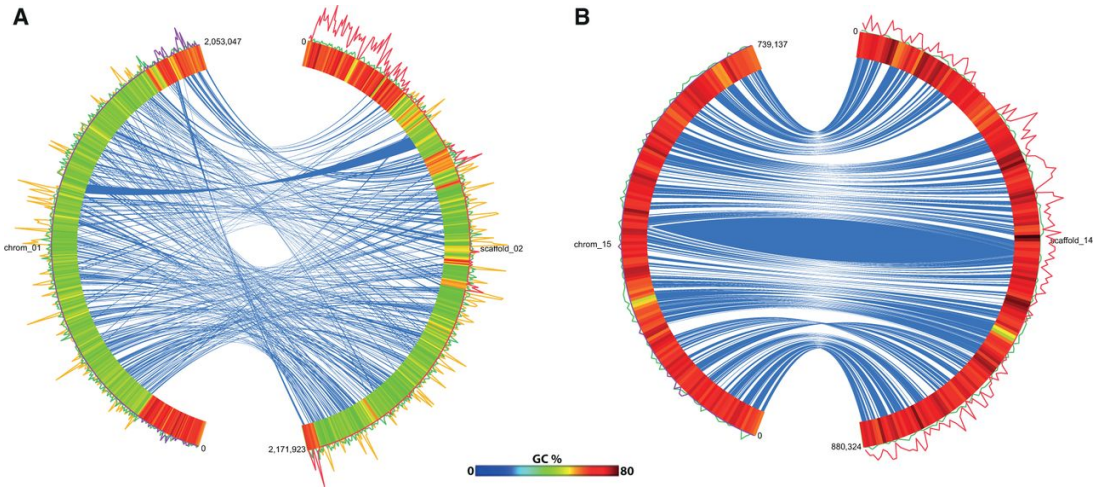


Figure 4.4. Genomic location of Introner Elements. (a) Comparison of BOC chromosomes of RCC299 (left; chrom_01) and CCMP1545 (right: scaffold_02). **(b)** Comparison of chromosome 15 of RCC299 and scaffold_14 of CCMP1545. The outer band represents the GC percentage across the chromosome, while the inner connections (blue) represent orthologous genes between the two strains. Intron density is displayed on the outside of the outer band: IE-A/IE-B/IE-D (red), IE-C (purple), canonical introns (green), and BOC1 introns (yellow). Plots were drawn using Circos [314].

We could not identify any sequence motif, both at the nucleotide level or the amino acid level that would correlate with the presence of IEs. There is also no insertion bias toward specific gene categories. On the other hand, the only functional category of genes completely lacking IEs involves genes that code for ribosomal structural components. However, it is well known that these genes are intron-poor and have a specific intron set – sometimes encoding small nucleolar RNAs – that helps to regulate the production and function of the ribosome [365], which could explain the absence of IEs due to strong selection against any further insertions.

The positioning of IEs within genes tends to favor the centre of the gene, which is similar to what has been recently reported for IE-like introns in fungi [286]. On the contrary, canonical introns in *Micromonas* are more often found at gene extremities *(Supplementary Figure 4.7)* and mostly in the genic 5' region [366], a feature primarily ascribed to intron loss at the genic 3' region [367].

4.2.4 Replication

When searching marine metagenomes at NCBI [http://www.ncbi.nlm.nih.gov, last accessed November 28, 2013] and CAMERA [368] for IEs, we uncovered 2,794 metagenomic sequences containing complete or partial IEs. This finding confirms that the IEs are not an artefactual strain feature but are present in the ocean within a wider variety of strain-related organisms. When comparing both *Micromonas* genomes to these metagenomic samples [368], we discovered Presence/Absence Polymorphisms (PAPs) of IEs *(Figure 4.5)*. In total, 913 metagenomic sequences revealed 511 unique novel IE insertions. Most metagenomic sequences containing IE-A elements were highly identical to the CCMP1545 genome, while for IE-C-containing sequences, a higher degree of diversity was found. At the same time, we discovered about 13 times more metagenomic sequences with novel IE-C positions compared with IE-A or IE-B/IE-D for which we have no proof for 'novel' insertions (IE-A: 35; IE-B: 0; IE-C: 476; IE-D: 0).

The difference in PAPs can be explained by IE-C either being more active or IE-C being more widespread, or a combination of both. Besides in metagenomic sequences, an occurrence of IE-C-containing sequences was observed within the CCMP1764 strain (*Micromonas pusilla* clade-I), for which short-read sequences have been obtained. After assembling the CCMP1764 genome, we compared it with the RCC299 genome. Only 31 IE-C positions are conserved in both genomes, while 149 and 66 are unique to RCC299 and CCMP1764, respectively, indicating that IE-C has been actively replicating since the divergence of RCC299 and CCMP1764.

Comparison with metagenomic data thus suggests that IEs are mobile elements that can replicate themselves

Figure 4.5. Presence/Absence Polymorphisms in *Micromonas*. Two alignments showing one IE-A1 and two IE-C PAPs (between scaffold_14 of CCMP1545 (position 384641-384976) and metagenomic read CAM_READ_0212050959, and between chromosome 6 of RCC299 (position 160647-160930) and metagenomic read CCMP1764_READ_00758094 respectively).

and transpose into new locations. IEs are only found in transcribed regions in the sense orientation, which suggests that their mobility is linked to the transcription/splicing process. The mechanism most likely to explain this scenario is known as intron transposition [253, 283] *(Figure 4.6)*. Under this scenario, an IE can invade a transcript by reverse splicing. The resulting IE-containing transcript is subsequently reverse transcribed after which the cDNA undergoes homologous recombination with the corresponding genomic locus. The final result is that the IE is now found at a novel position in the genomic sequence.

An analysis of orthologous introns between CCMP1545 and RCC299 genes revealed 32 cases of IE remnants buried within conserved canonical introns. There are also several cases of nested or merged IEs, i.e., IEs inserted inside or merged with another IE *(Supplementary Figure 4.8)*. Therefore, the 'mobility phase' of IEs has to occur at a stage that still features a non-spliced primary transcript and not at the mature mRNA level.

4.2.5 Complex intron landscape

The genomes of the tiny unicellular Mamiellophyceae are among the smallest found in eukaryotes [1, 173, 182, 221]. Genome analysis shows that they all lack the U12 minor spliceosome components [369]. Consequently, it is surprising to find such a complex intron landscape within this taxon, with *Micromonas* CCMP1545 harboring five different classes of U2 spliceosomal introns, a unique feature never documented in any other eukaryote up to now. Analysis of intron size in eukaryotic genomes usually gives a typical distribution, as shown for plants (using *Arabidopsis thaliana* as a representative) and algae (using *Chlamydomonas reinhardtii* as a representative) with a single or major peak of small introns and a tail or a shoulder of big introns, which usually results from the insertion of transposable elements or other repeat elements [362] *(Figure 4.1)*. Two of the five *Micromonas* intron classes, namely canonical introns and BOC1 introns, are observed in all species of Mamiellophyceae

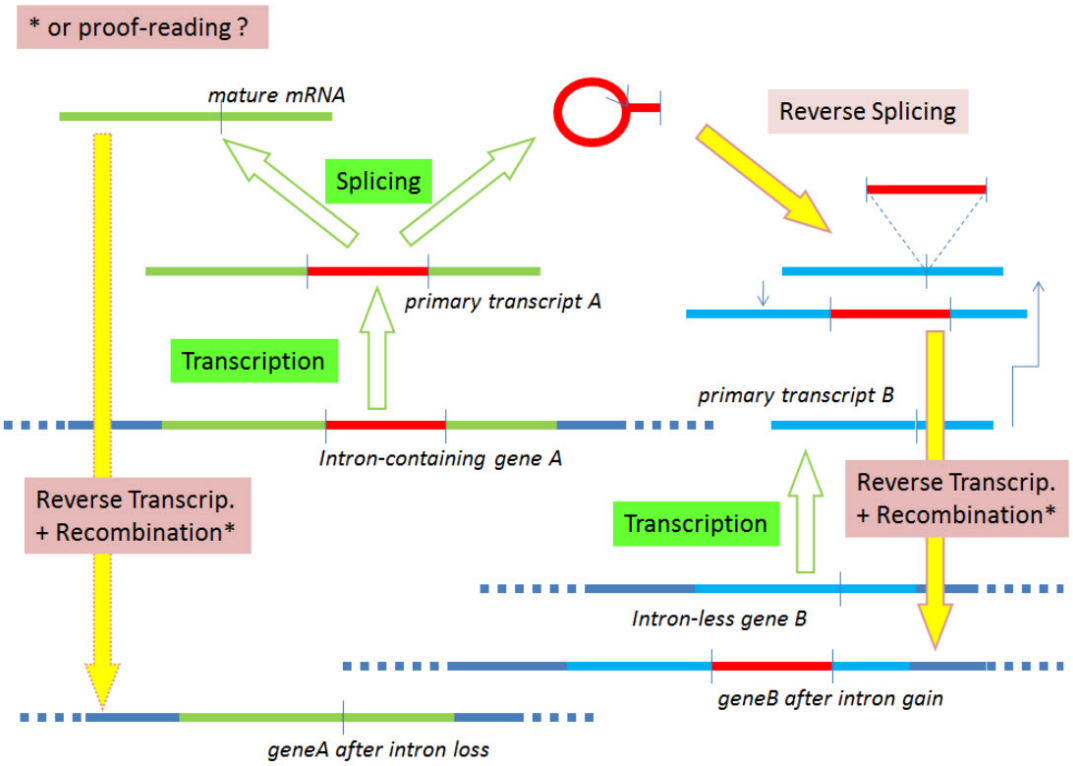


Figure 4.6. Introner Element replication mechanism.

[1]. Canonical introns are found on most chromosomes, contain conserved splice signals, and their number is limited to a few per gene. On the contrary, BOC1 introns are restricted to a specific area of the genome, do not display any conserved signals, and their hosting genes can contain high numbers of them. Adding to this complexity, we described the presence of four independent populations of invasive introns of unknown origin, with numbers amounting to some 6,100 copies in the CCMP1545 strain, compared with a population of 4,300 resident introns.

4.2.6 Intron evolution

The unique dual genome architecture of Mamiellophyceae, unicellular picoeukaryotes with an abundant population size, coupled with the extra complexity derived from the intron invasion, strongly contradicts the idea that intron-rich architecture complexity arose in multicellular eukaryotes of small population size [370, 371]. It is unclear how the U2 spliceosome is able to deal with the different intron classes that presumably have different splicing efficiencies, and which evolutionary mechanisms have directed this intron diversity and invasion. Since their discovery [372], the origin of spliceosomal introns in eukaryotes has been heavily debated, with tenants of the Introns-Early (InE) theory stating that the early eukaryotes already contained numerous introns, and proponents of the Introns-Late (InL) theory arguing for a gradual increase in intron numbers throughout evolution [373]. Among the latest proposals on the origin of spliceosomal introns, it was suggested that they were acquired from mitochondria group II introns at the dawn of eukaryote evolution, right after the engulfment of the bacterial ancestor giving rise to mitochondria. They would then have invaded the ancestral eukaryotic genome with a concomitant need to create a nuclear compartment that allows the slow process of splicing to be completed before translation could be initiated [265]. The presence of introns at homologous positions in orthologous genes in a large number of widely divergent eukaryotes rules in favor of the InE scenario, which consequently has led to the consensus that the Last Eukaryotic Common Ancestor contained intron-rich genes that more or less have been lost in different lineages [253, 374–376].

However, recent studies seem to imply that intron gain is more widespread than previously thought [377], leading to a more balanced view of intron origin [378]. Recurrent intron gain in genes of prokaryotic origin has

been observed after lateral gene transfer to eukaryotic taxa, an event that was suggested to be selected in intron-rich host genomes by nonsense-mediated mRNA decay (NMD) [379].

Peculiar intron gains were recently observed in the pelagic tunicate *Oikopleura dioica* [361], the microcrustacean *Daphnia pulex* [295], the dothideomycete fungi *Mycosphaerella graminicola* [380] and *Cladosporium fulvum* [286], and of course *M. pusilla* CCMP1545 [221], a list to which we are now adding *Micromonas* sp. RCC299. Within the same species, newly gained introns were found to be highly similar in sequence, except for *Daphnia*. In *D. pulex*, 24 cases of intron gain were observed when comparing genomic sequences of two different genomes and sequences from natural isolates, but those gains were independent from each other, even gains occurring at the very same site. Regarding *O. dioica*, although its introns have several features in common with those of *Micromonas* – they are present mostly at unique positions and show noncanonical splice sites especially for newly gained introns – only four pairs of Nearly Identical Introns (NIIs) were found out of a total of ~75,000 introns. In this case, both NIIs in a pair were found within the same gene and were suggested to be the result of reverse splicing. In fungi, intron gain due to the insertion of NIIs (Introner-like Elements (ILEs), analogous to *Micromonas*' IEs), shares some features with IE insertions. Depending on the species, ILEs occur in a range of a few tens up to ~500, out of a total of more than 10,000 introns. Within the Mycosphaerellaceae species, they are related to each other, suggesting the presence of ILEs predating speciation within this clade ~100 Mya. ILEs were shown to be efficiently spliced but to share specific features compared with resident introns, such as a bigger size and a conserved secondary structure. Finally, ILEs were shown to slowly degenerate with time, loosing progressively these specific features, and were thus suggested to be ancestors of many resident introns.

What makes *Micromonas* stand out is first and foremost the amplitude of intron gain, with hundreds to thousands of newly gained introns, comparable in number to an invasion of transposable elements. Because of its huge numbers, IE invasion can truly be seen as an Introns-Late case, in which the organisms' intron content is significantly enriched, more than doubled in the case of CCMP1545. These IE numbers must impact the biology of *Micromonas*, while the other reported intron gains would likely not. The second difference lies in the genome characteristics. *Micromonas*, just like all other Mamiellophyceae, only contains a few resident introns, whereas the organisms listed above are intron-rich, although to a lower extent for Mycosphaerellaceae fungi, for which the number of introns lies between 1 and 2 introns per gene [381]. The argument of intron gain as a way to homogenize gene architecture through NMD [370, 371] is falling short with the *Micromonas* IEs. Contrary to ILEs, we did not observe a clear or peculiar secondary structure within IEs. Finally, the intron invasion in the unicellular *Micromonas* goes against 'simple population-genetic principles' stating that the selective disadvantage of intron-containing alleles, even if weak, would be a barrier to the proliferation of introns in organisms with a huge population size [382].

We propose that, at a given point during evolution, a genetic element such as the IE has arisen after which it started to replicate, as for ILEs. Because all intron gain events listed above vary greatly in sequence, these events must have happened independently from each other, in contrast to ILEs. In the case of both *Micromonas* isolates, metagenomic evidence suggests that IE-C is present in a wider variety of host *Micromonas* organisms, as metagenomic sequences containing IE-C display a higher degree of sequence variety than IE-A/IE-B ones. This explains why *M. pusilla* CCMP1764, which belongs to a different clade than RCC299 [221], also carries IE-C sequences. IE-C therefore needs to have originated in an ancestor of clade I and II, but after the divergence of clades III and V. As of now, IE-A/IE-B seems to be restricted to clade V *(Figure 4.7)*.

As reported for fungi [286], IEs degrade over time and undergo mutations and indels (with a bias toward deletions) until the IE signature 'fades out'. It is therefore possible that many of the *Micromonas* introns that we now label as canonical are in fact highly degraded IEs.

Various mechanisms have been suggested to explain intron gains, such as gene duplication, insertion of transposable elements, mutational creation of novel splice sites, or splicing enhancing features. Our findings as well as other recent ones implying propagation of intron copies do favor the reverse-splicing/recombination scenario [282] suggested earlier by Cavalier-Smith [383]. In the first step of this scenario, an intron freed from one pre-mRNA would be inserted into another pre-mRNA by the splicing machinery *(Figure 4.6)*. Reverse splicing, which was initially a rather wild hypothesis, nowadays turns out to fit with the current knowledge as it has recently been established in yeast that the two splicing steps are indeed reversible [284]. The second

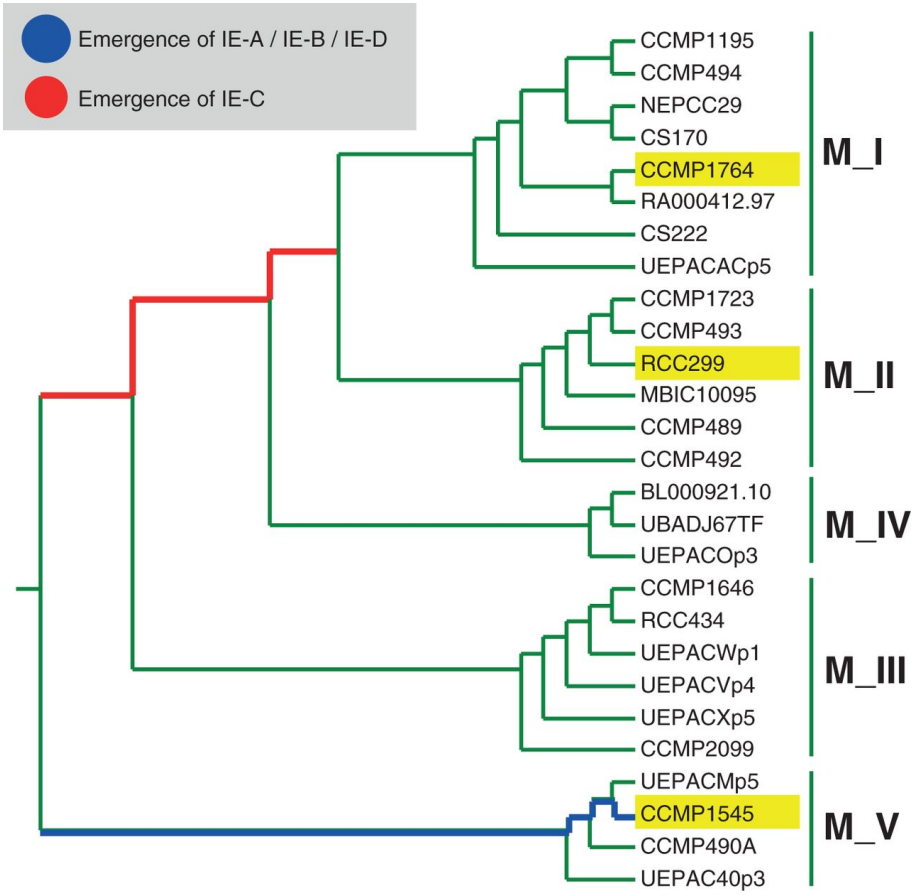


Figure 4.7. *Micromonas* phylogeny (adapted from Worden *et al.* [221]) inferred by neighbour joining based on 18S rRNA sequences. The tree shows the time windows in which the different IE classes have likely emerged with relation to the divergence of *Micromonas* isolates and strains and their clustering into five major clades, M_I to M_V. Isolates mentioned in the article are highlighted (yellow).

and third step should be the retro-transcription of the pre-mRNA into cDNA and the subsequent homologous recombination of this cDNA with its genomic partner *(Figure 4.6)*, both steps being documented in model eukaryotes and supported by the occurrence of intron loss for which they are required as well.

Why are IEs and other copy-introns specifically invasive and which features make these introns so successful in their capability to invade genomes while resident introns are generally noninvasive? Analysis of the transcriptome shows that transcripts for IE-containing genes are often not properly spliced, with many copies showing intron retention of IEs. This observation, together with the unusually high occurrence of noncanonical splice sites, argues for the *Micromonas* spliceosome to be permissive but rather ineffective for the newcomer introns that have not yet evolved the most efficient splicing mechanism, a hypothesis previously been put forward to explain evolution of mechanisms of RNA surveillance [384]. As a consequence, one would expect that IE splicing inefficiency would end up in promoting a proofreading mechanism, shunting the refractory spliceosome-bound pre-mRNA to a discard pathway [385]. This alone may in turn increase the chance for IE reverse splicing, which has been experimentally shown to happen under circumstances that favor spliceosome trans-conformation [284].

Some questions remain. What is the pace at which mobile introns are created and how long do they remain invasive? Are the mechanisms that control intron abundance similar to those observed for transposable elements? Finally, are *Micromonas* IEs and other cases of mobile introns just isolated exceptions to the rule, or are we on the verge of discovering many more hidden cases which would impact our view on the evolution of eukaryotic genome architecture, where intron invasion in eukaryotes would have occurred continuously?

4.3 Conclusions

The *Micromonas* strains CCMP1545 and RCC299 display a complex intron landscape, carrying canonical spliceosomal introns, Mamiellophyceae-specific introns (BOC1), and different classes of IEs. These IEs have colonized the genome by copying themselves into genes, likely involving reverse splicing. The findings presented in this article further strengthen the idea that intron gain is more widespread than previously thought.

4.4 Materials and methods

4.4.1 Sequence data

Micromonas genome sequences (v2.0) as well as the Expressed Sequence Tag (EST) libraries were obtained from the JGI portal [386, last accessed November 28, 2013]. Metagenomic sequences containing IEs were obtained (through BlastN) from the NCBI metagenomes database (taxid: 408169) and the CAMERA portal [368] using a handpicked set of ten IE sequences as query input. The CCMP1764 genome draft was assembled from the CAMERA CCMP1764 project data using the CLC Assembly Cell (v4.0.10; -b 110 -w 64).

Arabidopsis thaliana intron data were obtained from the TAIR10 intron database (v20101028), while *C. reinhardtii* intron data were derived from the latest Phytozome release (v5.3.1). When multiple isoforms were present, one representative was selected randomly.

4.4.2 IE prediction

IEs were predicted using a pattern matching approach, complemented with protein and EST evidence. Starting from handpicked example Introner Elements (IEs), we delineated common motifs (pattern blocks) and assembled them into class-specific pattern files (IEA-1 example listed below). We used PatScan [387] to scan the *Micromonas* genomes. For each class, multiple pattern files were constructed, ranging from strict to degenerate. When overlapping matches were detected, only the match belonging to the strictest pattern file was kept. EST and protein alignments were generated using GenomeThreader [347] (v1.4.6; -minalignmentscore 0.95 -mincoverage 0.89) and the splicing information was used in the automated curation of the final set of predicted IEs i.e. adjusting IE start and stop coordinates to match the exon/intron boundaries.

```
p1=GTGCGT
0...15
p2=ACTGGTYCCCRTACGACC[5,0,0]
0...80
p3=STTTCAAT[2,0,0]
0...40
p4=GCCTTTCAACTC[3,0,0]
0...100
p5=AG
```

IE remnants were detected using BLASTN (v2.2.17; -e 1e-05)[388] using the previously built set of (complete) IEs. For each class, we also built a multiple sequence alignment (MSA), constructed a profile HMM (HMMer v2.3.2), and used it to detect additional instances of degenerated IEs. This HMM approach was also used for members of the IE-B / IE-D class.

4.4.3 Reannotation of *Micromonas* genomes

Gene models were extensively curated through automated and manual procedures. All intron and gene information is stored in a relational database and can be accessed through the ORCAE platform (last accessed November 28, 2013) [110, http://bioinformatics.psb.ugent.be/orcae]. Data sets (gene models, intron sets, and environmental sequences) can be obtained from its download section.

4.4.4 *Micromonas* intron classification: BOC1 and canonical introns

BOC1 introns are defined as short (<75 nt), AT-rich (<43 GC%) introns lying in the BOC1 region of chromosome 1 of CCMP1545 (position 438,300-2,118,000) and chromosome 2 of RCC299 (position 263,000-1,817,000) [1]. Canonical introns are defined as all remaining introns that do not fall in either the IE or BOC1 categories.

4.4.5 Orthologous *Micromonas* introns

In total, 6,891 one-to-one orthologous pairs were identified using orthoMCL (v2.0; mcl options: -abc -I 1.5), representing 74% of the total intron content of both *Micromonas* isolates. After alignment (MUSCLE v3.8.31; -diags), intron positions were compared and cross-referenced against their class identifier (IE-A, IE-B, IE-C, BOC1, canonical).

4.4.6 Gene ontology analysis of IE genes

GO terms for all *Micromonas* proteins were derived using InterPro2GO [126], and GO term over/under-representation of genes carrying IEs, using the GO terms of the entire *Micromonas* proteome as a background, was analysed using the Cytoscape plugin BiNGO [389] (hypergeometric test + FDR correction; significance level 0.05). This GO analysis was only performed on CCMP1545, as the low number of IEs in RCC299 makes the analysis insignificant.

4.4.7 Spliceosomal components

Spliceosomal components were detected through homology with *A. thaliana* proteins in the Splicing Related Gene Database (`http://www.plantgdb.org/SRGD/`) and through the detection of splicing-related GO labels.

4.4.8 Metagenomic analysis

Metagenomic sequences were subjected to the IE prediction pipeline and aligned to the *Micromonas* genomes using a seed-and-align procedure, initiated by a regular BLASTN search. Starting from the best-hit, we expanded the genomic space with neighbouring hits. In the end, we used the outer coordinates to extract the corresponding genomic region, and re-aligned it to the environmental sequence using a SMITH-WATERMAN alignment (EMBOSS [390]: water).

To be able to draw accurate conclusions on IE Presence/Absence Polymorphisms (PAPs), we performed a quality filtering step. We only continued with alignments that have more than 100 nucleotides labelled as 'non-IE', an identity percentage of more than 50%, and a coverage of more than 60%. After careful consideration, we also decided to leave aside all metagenomic sequences labelled as 'JCVI', as they were assemblies of smaller metagenomic sequences. After this quality filtering, we then compared IE positions to PAPs. This analysis is highly biased toward the finding of IEs that are absent in RCC299/CCMP1545 but present in the metagenomic sequences, as the reverse would require a confirmation that the read is derived from an organism that carries the specific IE. This is only the case when a sequence carries a strain identifier (i.e., as with the CCMP1764 case) or if the metagenomic sequence carries an IE up- or downstream.

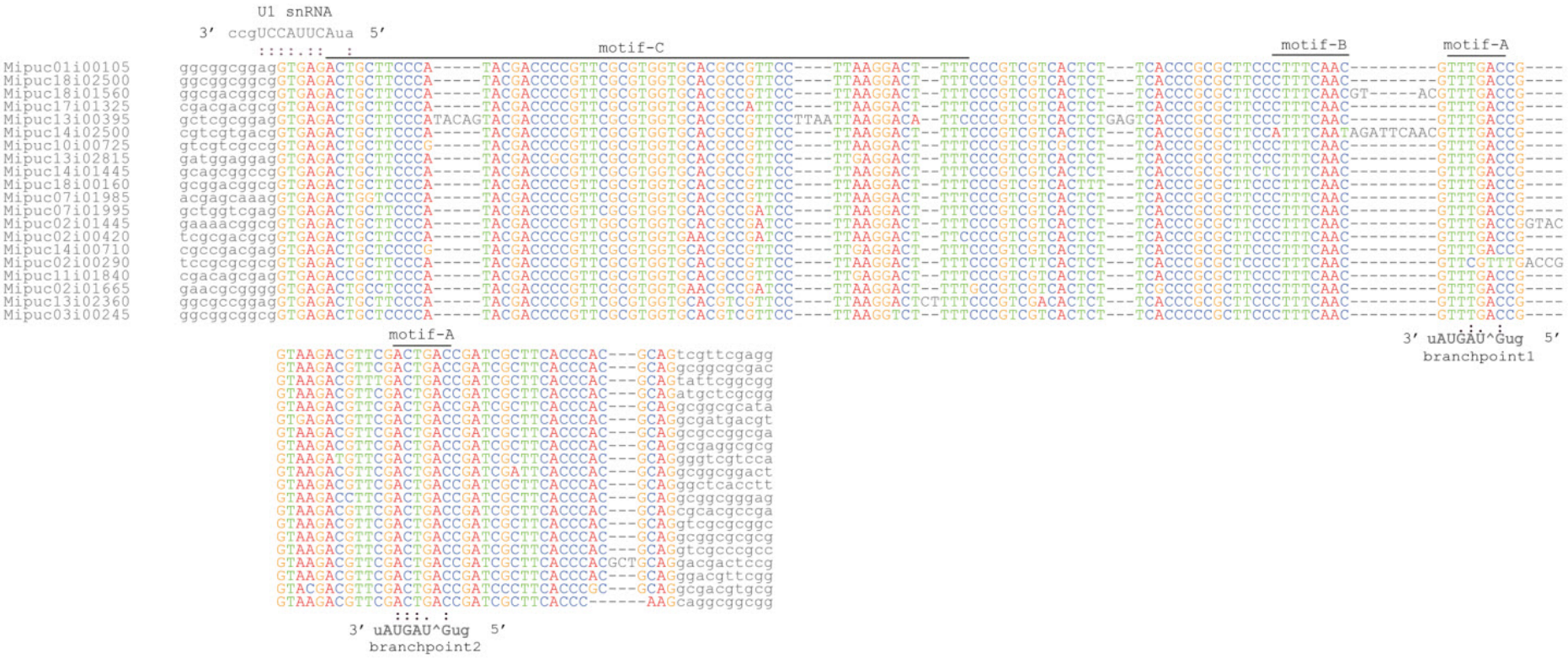
4.5 Supplementary Information

This section contains selected segments of supplementary figures most relevant to this chapter and the topic of this dissertation.

Supplementary Figure 4.1. *Micromonas* splice site signals for all intron classes. Shown here are sequence logos for the donor/aceptor site (10 nucleotides upstream and downwstream) for CCMP1545 **(a)** and RCC299 **(b)**.

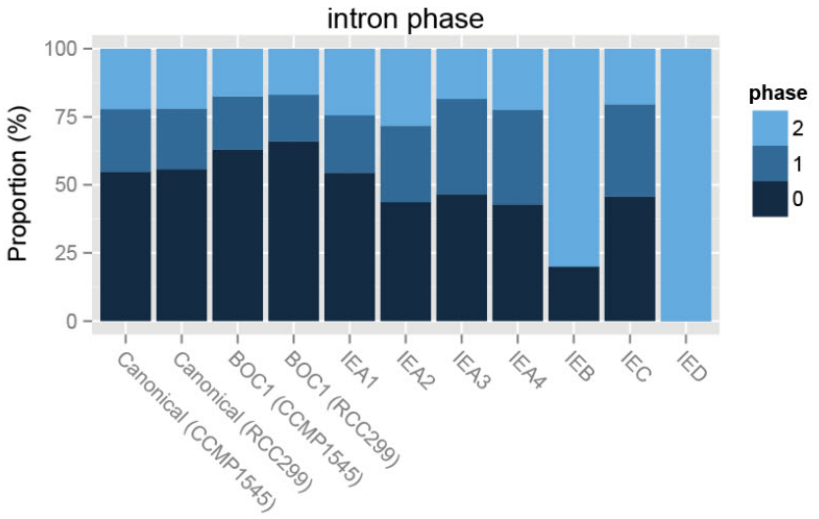
Supplementary Figure 4.2. Alignment of 20 random IE-A1 sequences. The motifs (motif-A = branchpoint motif; motif-B = branch-point companion motif; motif-C) are marked, as are the splicing signals (donor site + branchpoint) and their base-pairing information from the corresponding spliceosomal RNAs.

Supplementary Figure 4.3. Alignment of 20 random IE-A2 sequences. The motifs (motif-A = branchpoint motif; motif-B = branch-point companion motif; motif-C) are marked, as are the splicing signals (donor site + branchpoint) and their base-pairing information from the corresponding spliceosomal RNAs.

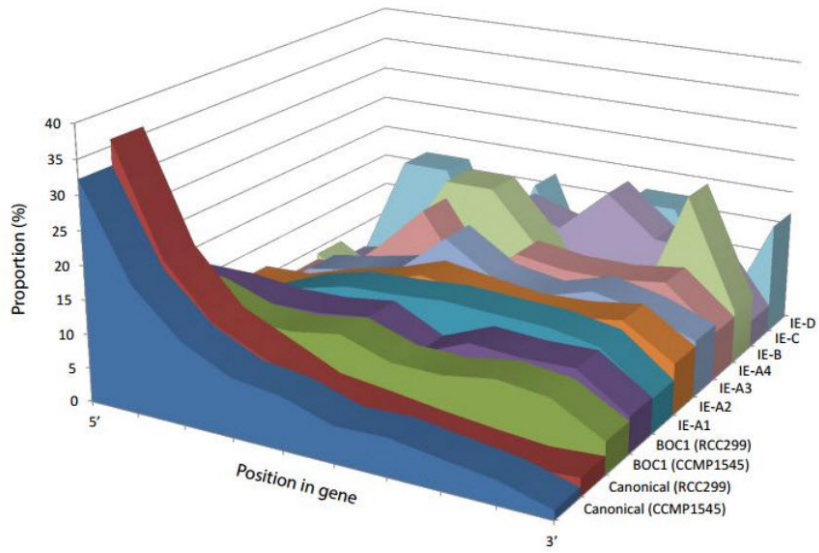


Supplementary Figure 4.4. Alignment of 20 random IE-A3 sequences. The motifs (motif-A = branchpoint motif; motif-B = branch-point companion motif; motif-C) are marked, as are the splicing signals (donor site + branchpoint) and their base-pairing information from the corresponding spliceosomal RNAs.

Supplementary Figure 4.5. Alignment of 20 random IE-A4 sequences. The motifs (motif-A = branchpoint motif; motif-B = branch-point companion motif; motif-C) are marked, as are the splicing signals (donor site + branchpoint) and their base-pairing information from the corresponding spliceosomal RNAs.



Supplementary Figure 4.6. Phase distribution of Introner Elements, BOC1 and canonical introns.



Supplementary Figure 4.7. Positioning of Introner Elements, BOC1 and canonical introns inside genes.

Supplementary Figure 4.8. Merged Introner Elements. Alignment between *Micromonas pusilla* CCMP1545 scaffold_14 (position 341989-343470) and metagenomic read AACY02323272 (GenBank Accession Number). Due to loss of internal splice structures when merging, it is hard to exactly delineate borders.

CHAPTER 5

AN IMPROVED GENOME OF THE MODEL MARINE ALGA *OSTREOCOCCUS TAURI* UNFOLDS BY ASSESSING ILLUMINA *DE NOVO* ASSEMBLIES

*Romain Blanc-Mathieu, **Bram Verhelst**, Evelyne Derelle, Stephane Rombouts, François Yves-Bouget, Isabelle Carré, Annie Château, Adam Eyre-Walker, Nigel Grimsley, Hervé Moreau, Benoit Piégu, Eric Rivals, Wendy Schackwitz, Yves Van de Peer and Gwenaël Piganeau*

Abstract

Cost effective Next-Generation Sequencing technologies now enable the production of genomic datasets for many novel planktonic eukaryotes, representing an understudied reservoir of genetic diversity. *O. tauri* is the smallest free-living photosynthetic eukaryote known to date, a coccoid green alga that was first isolated in 1995 in a lagoon by the Mediterranean sea. Its simple features, ease of culture and the sequencing of its 13 Mb haploid nuclear genome have promoted this microalga as a new model organism for cell biology. Here, we investigated the quality of genome assemblies of Illumina GAIIx 75 bp paired-end reads from *Ostreococcus tauri*, thereby also improving the existing assembly and showing the genome to be stably maintained in culture.

The 3 assemblers used, ABySS, CLCBio and Velvet, produced 95% complete genomes in 1402 to 2080 scaffolds with a very low rate of misassembly. Reciprocally, these assemblies improved the original genome assembly by filling in 930 gaps. Combined with additional analysis of raw reads and PCR sequencing effort, 1194 gaps have been solved in total adding up to 460 kb of sequence. Mapping of RNAseq Illumina data on this updated genome led to a twofold reduction in the proportion of multi-exon protein coding genes, representing 19% of the total 7699 protein coding genes. The comparison of the DNA extracted in 2001 and 2009 revealed the fixation of 8 single nucleotide substitutions and 2 deletions during the approximately 6000 generations in the lab. The deletions either knocked out or truncated two predicted transmembrane proteins, including a glutamate-receptor like gene.

Contributions

- Structural and functional annotation
- Manual gene curation
- RNAseq mapping and annotation quality assessment
- Maintenance of the *Ostreococcus*-section on the ORCAE platform
- Writing manuscript segments (in respect to the topics mentioned above)

5.1 Introduction

Unicellular marine photosynthetic eukaryotic organisms represent much of the untapped genetic diversity reservoir of our planet [391, 392]. Their ecological importance in the global carbon cycle [298, 393] and their biotechnological potential as possible sources of biofuels and dietary «omega-3» lipid food supplements, have fostered several genome projects to gain knowledge into their diversity and metabolic potential [1, 181, 182, 221, 394, 395]. *Ostreococcus tauri* is the smallest photosynthetic eukaryote known and its genome was the first marine green algal genome to be sequenced. It has a simple cellular organization with a single mitochondrion and a single chloroplast [170, 171], all orchestrated by a 13 Mb haploid nuclear genome [182]. Its compact genome, ease of culture and genetic transformation by homologous recombination promoted *O. tauri* as an ideal model for cell biology [190, 396]. It has been successfully used to gain knowledge into fundamental cellular processes such as the cell cycle [194, 199, 397], the circadian clock [202, 205, 398], lipid [399] and starch synthesis [400], as well as the mechanisms of genome evolution [318, 401, 402].

High throughput technologies approaches are revolutionizing research on phytoplanktonic eukaryotes [403]. Illumina, among the market leaders for low cost nucleotide sequencing [404], has been broadly adopted for sequencing phytoplanktonic eukaryotes. To what extent this approach delivers worthy genome sequence therefore merits critical appraisal. Comparative studies to assess the quality of *de novo* assemblies are scarce and suggest that assembly quality varies widely from one species to another and from one assembler to another [150]. Even fewer studies have been made to evaluate the quality and accuracy of *de novo* scaffolds [405], as the major limiting step is the availability of a high quality reference genome sequence to benchmark an assembly resulting from processing short reads.

DNA was extracted from the *O. tauri* strain in 2001 (OT-2001) and in 2009 (OT-2009) and 40 millions paired-end DNAseq reads were generated from each extraction. These datasets were used to compare the output of three *de novo* assembly algorithms. The resulting assemblies were benchmarked against the *O. tauri* sequenced genome to estimate their quality and the percent of the genome covered. Combined with RNAseq data, this data led to a significant improvement of the reference genome sequence by resolving 1194 gaps corresponding to 460 kb and resulting in a remarkable improvement of the 7699 protein coding genes models.

Genetic selection pressures differ between organisms that grow in the wild, that are subject for example to limiting environmental conditions (such as nutrient supply) and in the laboratory, where mutations favouring growth in culture are expected to become fixed over time [406]. Previous studies on a few genes have revealed amino-acid differences that result in marked differences in the phenotype of the *S. cereviseae* lab strain as compared to wild strains [407, 408]. More recently, Illumina sequencing allowed scientists to track 120 mutations in yeast during three experiments selecting for increased growth rates in a constant environment [409]. The *O. tauri* strain has been maintained in laboratory culture conditions since its isolation in 1995 [170]. The comparison between the 2001 and 2009 sequence data enabled us to investigate genome stability of *O. tauri* over approximately 6000 generations of lab subculturing.

5.2 Results and discussion

5.2.1 *De novo* assemblies of *O. tauri*'s genome

We generated *de novo* assemblies of the *O. tauri* genome using 41 million paired-end 76 bp Illumina reads from the OT-2009 strain. The three different assembly algorithms produced between 1402 to 2080 scaffolds, with a weighted median size length (N50) of 9,539 to 14,550 bp *(Table 5.1)*. The total assembly size varied from 12.3 to 12.8 Mb and corresponds to 94 to 96% of the complete Genbank reference genome sequence. Among the three assemblies, ABySS and CLCbio produced assemblies with better contiguity; they had fewer scaffolds (CLC: 1402 and Ab: 1490) with greater N50 (Ab 14550, CLC 14519) and both covered 96% of the Genbank reference genome sequence. ABySS produced the longest assembly of 12.8 Mb, closest to the expected *O. tauri* genome size. The alignment of the *de novo* assembly generated by ABySS on to the *O. tauri* reference genome sequence is presented in *Figure 5.1* (outer circle).

Assembly (kmer_size)	Nb of scaffold	N50	Size (Mb)	Nb of Aligned Scaffolds	Aligned Bases (Mb)	Ref. cov.[1] (%)	CDS cov.[2] (%)	Start to Stop CDS[3] (%)
Velvet (41)	2080	959	12.3	2066	11.68	94	96	42
ABySS (31)	1490	14550	12.8	1474	11.87	95	98	43
CLCbio (28)	1402	14519	12.6	1394	11.96	96	98	42

Table 5.1. Assembly statistics of *de novo* assemblers in *O. tauri*. [1,2]percentage of aligned bases against the reference genome sequence and against the coding sequences (CDSs). [3]percentage of complete CDS within a single scaffold.

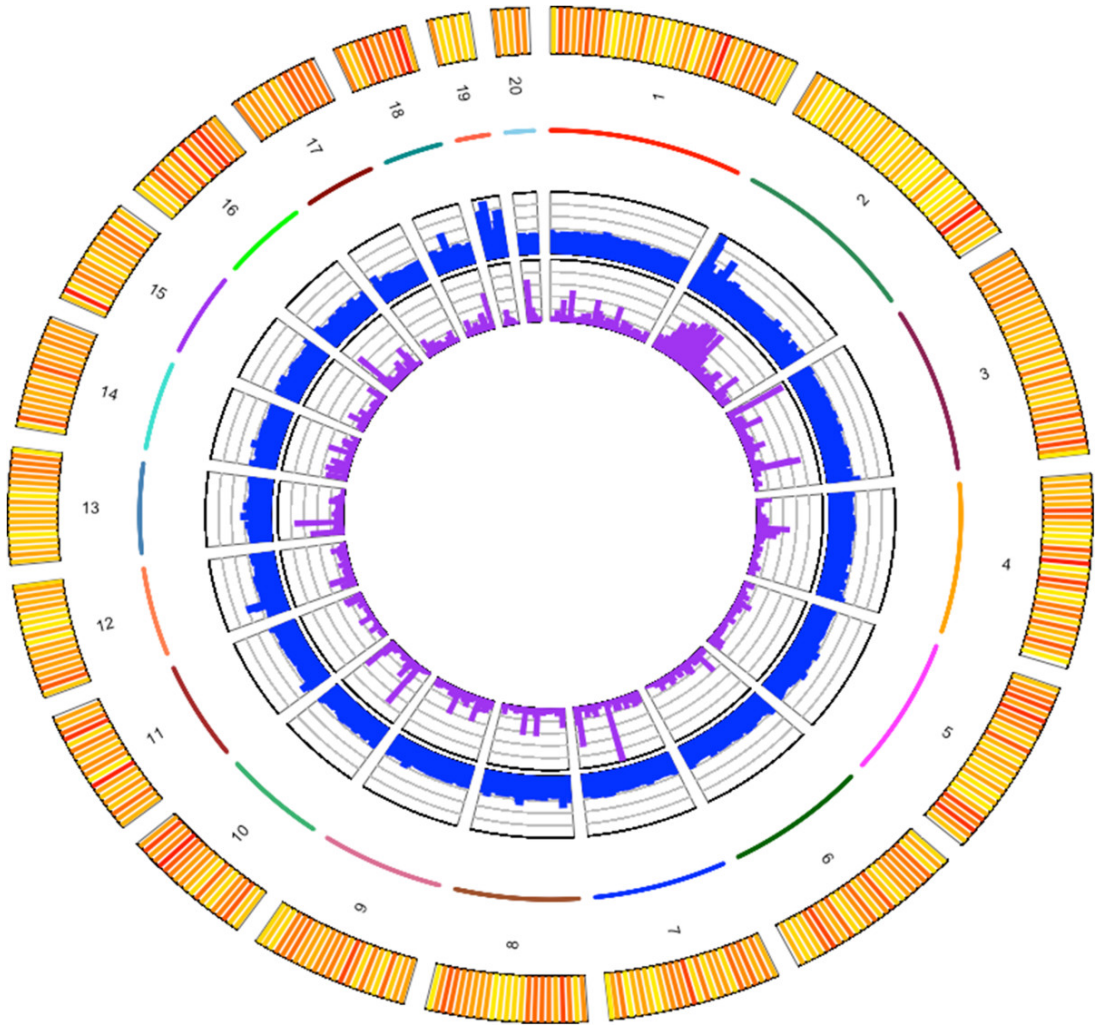


Figure 5.1. Illumina DNAseq and RNAseq aligned against *Ostreococcus tauri* reference genome sequence. Colored numbered lines represent the 20 chromosomes of *Ostreococcus tauri*. The contiguity of the *de novo* assembly along the chromosomes ranges from 0 (white) to 28 scaffolds per 30 kb window (red). The inner blue track is the DNAseq coverage (from 0 to 582 reads per bp). The inner purple track is the RNAseq coverage averaged across 10 kb windows (from 0 to 1947 reads per bp). Figure generated with the RCircos software [410].

It is essential to assess the correctness of *de novo* assemblies as contiguity may come with a trade off in correctness [150]. dnadiff tools from the NUCmer alignment of each set of *de novo* scaffolds detected 5, 8 and 6 translocations, 3, 4 and 7 relocations for Velvet, ABySS and CLCbio respectively. The average number of mis-joins per scaffolds was less than 0.009 and the percentage of mis-assembled scaffolds was less than 0.9 percent for the three assemblers *(Table 5.2)*.

Assembly	Misjoin			Mis-assembled	Average
	Translocation	Relocation	Inversion	scaffold (%)	misjoins/scaffold
Velvet	5	3	0	0.4	0.004
ABySS	8	4	0	0.9	0.009
CLCbio	6	7	0	0.9	0.009

Table 5.2. Correctness Statistics of each assembly assessed with dnadiff.

The coding sequence representation in these assemblies, measured as the percentage of coding sequence base pairs in the original assembly that align against a *de novo* scaffold is 96.1, 97.6 and 98.2 (Velvet, ABySS and CLCbio) *(Table 5.2)*. This is significantly higher than that observed for intergenic regions (86.8, 84.9, 87.2 for ABySS, Velvet and CLCbio respectively, Fisher exact test: p-value < 2.2×10^{-16} for all 3 assemblers). The number of coding sequences (CDSs) included from start to stop codon within a scaffold was 3101 (41.5%) for Velvet, 3363 (42.7%) for ABySS and 3274 (41.5%) for CLCbio.

To estimate the impact of sequencing depth on reference genome coverage and *de novo* assembly, we randomly sampled paired-end reads from our dataset to produce seven subsets corresponding to a 10, 25, 80, 125, 200, 225 and 250 fold sequencing depth. The different sampled paired-end reads sets were aligned against the reference genome using BWA and reassembled *de novo*. The obtained scaffolds were aligned against the reference genome using NUCmer. Figure 5.2 shows the relationship between raw reads and scaffolds genome coverage, and sequencing depth. Coverage changed from 99.5% to 94.8% when the sequencing depth decreased from 250X to 10X. It decreased more dramatically for *de novo* assembly, from 95% for sequencing depth greater or equal to 80X, down to 69% for a sequencing depth of 10X. This suggests that 80 fold sequencing depth is optimal for *de novo* genome assembly with this approach.

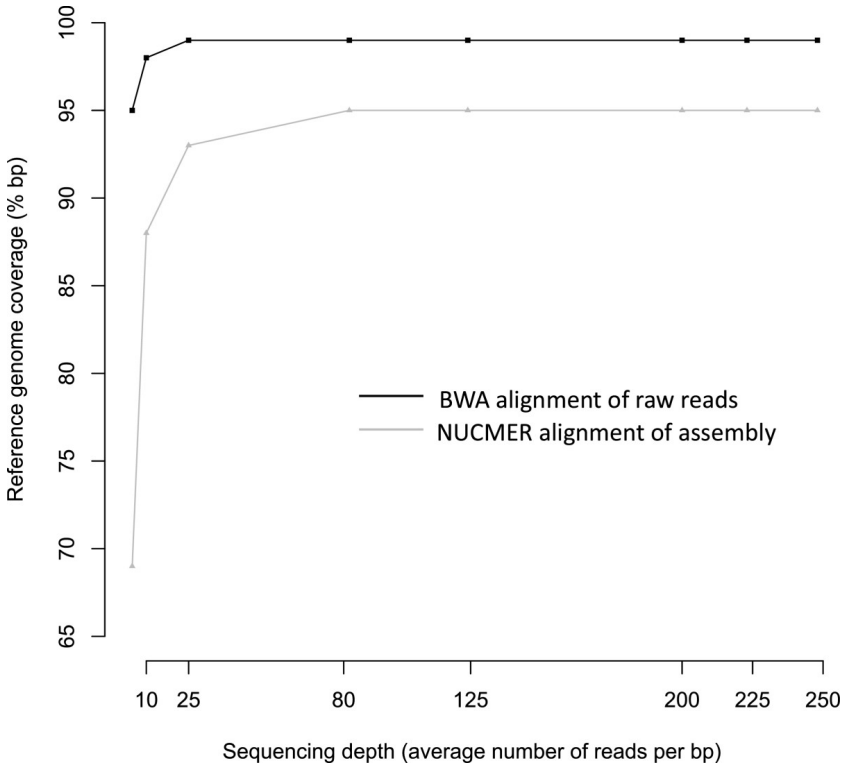


Figure 5.2. Saturation curve of coverage along the GenBank reference genome sequence. BWA alignment of 41 M Illumina paired-end reads subsets representing different sequencing depth (black line) and after NUCmer alignment of *de novo* scaffolds produced by a Velvet *de novo* assembly of these same paired-end reads subset (grey line).

5.2.2 Improving a historical genome sequence

The reference genome contained 1671 gaps, of which 930 could be resolved using *de novo* assembly and 92 could be further resolved by IMAGE [411]. In depth analysis of the remaining reads using CRAC identified 50 adjacent contigs linked by paired reads, of which 34 could be fixed, while two indicated clear assembly errors in the reference genome. These errors consisted of two inversions in the reference assembly. Fixing these two inversions closed 4 additional gaps. Additional 134 gaps were filled by PCR sequencing effort. The analysis of the alignment of the raw reads onto the updated genome sequence confirmed that the 477 still remaining gaps could not be joined by paired-end reads, as expected if they correspond to regions larger than 100 bp, or if the Illumina library did not contain the corresponding sequence. The 477 remaining gaps have a random distribution across the chromosomes (the distribution of the distances between gaps is not significantly different from expectations, Chi2 test, p = 0.43). The updated genome sequence is thus 12,916,858 nucleotides long, 460.5 kb longer than the historical reference genome sequence [182]. Alignment of paired-end reads against the reference genome sequence enabled 2126 single nucleotides and 3342 indel differences to be identified and corrected in the updated genome sequence.

5.2.3 Genome evolution between 2001 and 2009

Comparison of the OT-2001 and OT-2009 datasets enabled us to identify 8 nucleotide substitutions, 2 deletions and 1 insertion that had occurred in this strain between the 2001 and 2009 cultures *(Table 5.3)*. All except the insertion were confirmed by independent Sanger sequencing on the OT-2001 DNA and the OT-2009 DNA. The predicted insertion in the first 145 bp of chromosome 9 could not be amplified because of its proximity to the telomere CCCTAAA repeats. One substitution is synonymous, 6 are non-synonymous and one corresponds to a nonsense mutation *(Table 5.3)*. In total, two substitutions result in the introduction of a stop codon in a coding sequence (the nonsense mutation and one of the deletions). This may lead to a gene knockout, or alternatively, cause a shorter protein by initiation of translation from a downstream methionine *(Figure 5.3)*. For both genes, the entire genomic region is covered by RNAseq data.The analysis of read coverage over 50 bp windows along the chromosomes led to the identification of two large duplicated regions encompassing 80 kb on chromosome 19 and 30 kb on chromosome 2 *(Figure 5.1, inner circle)*. Local peaks on chromosome 12 and 18 correspond to the region containing the rRNA operon (ch12) and a single gene with unknown function, ostta18g00700 (ch18). Using coverage to estimate the number of gene copies, we predicted that there are 4 copies of the ribosomal gene « operon » and 5 copies of the ostta18g00700 gene. However, there is no evidence for copy number variations between 2001 and 2009 as no coverage variations have been identified between OT-2001 and OT-2009.

Chrom	Position	2001	2009	Type	CDS	Annotation
Ch3	333101	T	C	Non-Syn	0t03g02090	Unknown
Ch3	829938	T	A	Non-Syn	Ot03g05020	Metal-dependent hydrolase
Ch5	180669	C	T	Syn	Ot05g01240	Transcription factor NF-X1
Ch5	224089	A	T	Non-Syn	Ot05g01550	Dehydrogenase
Ch6	28989	G	C	Nonsense	Ot06g00160	Unknown
Ch6	772097	G	A	Non-Syn	Ot06g04800	Dynein 1-alpha heavy chain
Ch12	137126	C	A	Non-Syn	Ot12g00990	Glutamate receptor-related
Ch12	137173	C	T	Non-Syn	Ot12g00990	Glutamate receptor-related
Ch12	137177	T	G del	Frameshift	Ot12g00990	Glutamate receptor-related
Ch17	13580	C	GTCCAT del	Deletion	Ot17g00070	Heat shock protein 90
Ch9	145	A	C ins	Insertion	non coding	Telomeric region

Table 5.3. Evolution of the Genome sequence between 2001 and 2009.

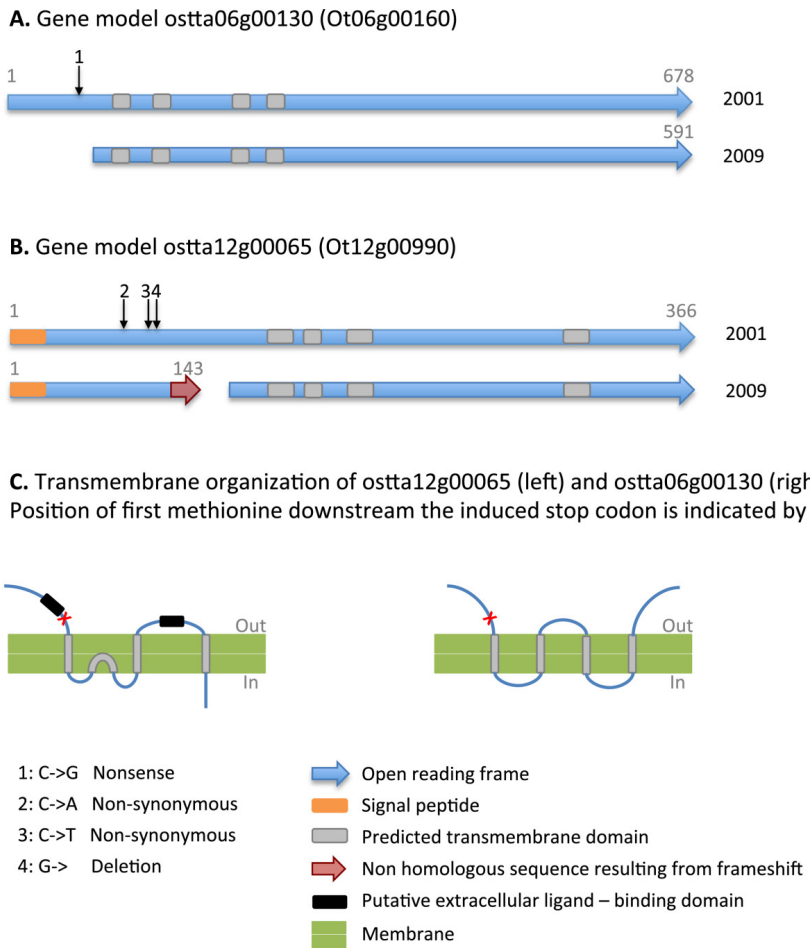


Figure 5.3. Localization of the substitutions between 2001 and 2009 within two genes. (a) Gene organization of ostta06g00130 (Ot06g00160) **(b)** Gene organization of ostta12g00065 (Ot12g00160) **(c)** Transmembrane organization of the two encoded proteins (left:configuration of *Arabidopsis* glutamate-like recepters, homologous to Ot12g00160 [412], left: TMHMM prediction for Ot06g00160.

5.2.4 Annotation update

Gene prediction from the updated genome sequence, followed by manual editing by experts, led to the annotation of 7699 protein coding genes, 39 tRNA and 319 transposable elements (TEs) *(Table 5.4)*. Compared to the annotation of the historical genome sequence, (protein-coding) genes are longer and contain fewer introns *(Table 5.4)*, while the proportion of validated introns - as measured by RNAseq - has risen drastically from 7% to 89%. The complete updated genome sequence has been submitted to Genbank and is available under accession numbers CAID01000001 to CAID01000020. Old gene model names are provided as synonyms in new gene models and the link between updated gene and the previous annotations can be browsed via ORCAE [110, `http://bioinformatics.psb.ugent.be/orcae/overview/OsttaV2`].

Version	Total size (Mb)	Nb CDS	Average gene length (bp)	Nb of genes with introns	Intron Size (Average)	Nb of TE
2006	12.5	7890	1290	3186 (39%)	103	417
2013	12.9	7699	1387	1440 (19%)	140	319

Table 5.4. Genome annotation update of *O. tauri*.

5.2.5 Sequences lacking in the assemblies

The comparison of the *de novo* assemblies with the reference genome enabled us to investigate the features of sequences absent from the *de novo* assemblies. These sequences tend to contain significantly more intergenic regions. This is in line with previous studies showing an increased coverage in exons [413, 414]. This may be due to the higher proportion of low complexity sequences in intergenic regions, as these produce assembly forks that stop the contig elongation in the assembler [415]. Another assembly-independent reason is the lack of reads in the library for these regions. The genome sequence with no read coverage had an average GC content of 80%, consistent with an underrepresentation of extreme GC sequences in Illumina sequencing data [416]. Reciprocally, *de novo* assemblies closed 930 gaps (56%) in the historical genome sequence, these resolved gaps had an average length of 386 bp.

5.2.6 Genome evolution under laboratory conditions between 2001 and 2009

O. tauri was isolated in 1995 from the Thau lagoon in the NW Mediterranean Sea and conserved in the lab since. When introduced into the lab, organisms may evolve as a consequence of selection for better growth and as a consequence of the loss of selective pressures that are present in the wild [406, 409] In this study, the comparison of the DNAseq data from 2001 and 2009 gave us an insight into the genome evolution of a lab-adapted strain. There is no evidence for copy number variations and our analysis revealed 8 substitutions, 2 deletions and possibly one insertion, suggesting a high level of genome stability within this timeframe, which corresponds to approximately 6000 generations. These substitutions occur within 8 protein coding genes and one intergenic region *(Table 5.3)*. Interestingly, 2 substitutions and one deletion occurred in the same gene (Ot12g00990) annotated as a membrane receptor related to the Glutamate-like receptor gene family (GLR). GLR are homologs of mammalian ionotropic glutamate receptors, glutamate-activated ion channels involved in rapid synaptic transmission. Their initial discovery in *Arabidopsis thaliana* raised intriguing questions about the physiological functions of neurotransmitter-gated channels in plants and provided an insight into why plants make chemicals that act on human brain [412]. The function and ligand of plant GLR is an intense area of research ([417] for a review) and they are hypothesized to be potential amino acid sensors. The deletion induced a frameshift and splits the gene into one 146 aa and one 380 aa open reading frames, thus shortening one of the ligand fixation regions predicted to be outside the cell *(Figure 5.3)*. In the second gene annotated as a membrane protein (Ot06g00160), the open reading frame was shortened from 678 to 591 amino acids. High throughput transcript analysis in *S. cerevisiae* suggests that 60% of genes have transcript isoforms, with several cases of downstream methionine initiation [418]. While we do not know the extent of transcriptional heterogeneity from isoform profiling in *O. tauri*, the substitutions we report here may have been either compensated by the initiation of the gene from a downstream methionine or may have caused a knock out of this gene. While Ot06g00160 has homologous genes in the two other *Ostreococcus* spp. genomes sequenced, the orthologous gene family of Ot12g00990 does not include any gene from the species *O. lucimarinus*, suggesting that this gene is dispensable if knocked out. Subculturing produces a bottleneck of 6×10^5 cells per subculture, a population size that should be sufficiently large to prevent the fixation of deleterious mutations as a consequence of drift, suggesting that these substitutions between the strains are either neutral or advantageous in the lab environment. Kvitek & Sherlock [409] have tracked mutations in one strain of *S. cerevisiae* evolving in a constant environment and provided evidence that many of the mutations led to the loss of signalling pathways that usually sense a changing environment. When these mutant cells were faced with uncertain environments, the mutations proved to be deleterious. Consistent with this, the knock-out of two transmembrane genes may lead to altered perception of environmental signals, but this is difficult to test experimentally without knowledge of the signalling pathways that might be affected.

5.3 Conclusions

Although the *de novo* assemblies are fragmented in nature, we show that less than 5% of the genome is lacking from any *de novo* assembly. We took advantage of this data to improve the reference genome sequence of this model marine alga significantly and we show that only 9 substitutions have occurred within 6000 generations of lab culture.

5.4 Materials and methods

5.4.1 Data

We used the *O. tauri* whole genome sequence as a reference (GenBank accession number CAID01000001 to CAID01000020), sequenced on two BAC and five shotgun libraries [182]. The scaffolding was improved by using information about the location of each contig in a BAC library hybridized to macroarrays [182], leading to 20 scaffolds representing a total of 12.56 Mb, corresponding to 20 chromosomes. The reference genome assembly contained 1671 gaps as a consequence of low coverage (7X).

The culture used for the reference genome sequence came from a natural sample of *O. tauri* isolated 1995 in the Thau Lagoon [170, 171] and maintained by serial subcultures using 50 ml plastic tissue culture flasks in 20 ml K medium at 20°C under 100 μE s^{-1} m^{-2} constant light in Banyuls sur mer. Every 2 to 3 weeks the cells reach a stationary phase (at a concentration of approximately 3×10^7 cells.ml^{-1}) and 20 μl (approx. 6×10^5 cells) is sub-cultured in fresh K media. This culture was cloned in 2005 on agar plate and the cloned culture was maintained in the lab.

DNA extraction was performed on the 2001 and the 2009 culture as previously described [182]. Genomic DNA of the *Ot* strain from 2001 (OT-2001), from the same extraction sample that was used for Sanger sequencing, and 2009 (OT-2009) was randomly sheared into ~250-bp fragments. The libraries created from these fragments were sequenced on an Illumina GAIIx system at the Joint Genome Institute. The sequencing experiment produced 43 millions and 41 millions 76 bp paired-end reads with an average insert size of 250 nucleotides. The alignment of these paired-end reads against the reference genome sequence (BWA version 0.6.1-r104 with default parameters [419]), produced an average coverage of 175 and 205 reads per reference base pair in OT-2001 and OT-2009, respectively. Both 2001 and 2009 cultures were non-axenic and contained bacteria, as judged from the presence of bacterial contigs in the assemblies [420]. As the OT-2009 dataset corresponded to a clonal strain, this dataset was used for analysis of *de novo* assemblers and genome update. The clonal strain resequenced in 2009 has been submitted to the Roscoff Culture Collection under accession number RCC4221. The Illumina dataset have been deposited in the SRA archive under accession numbers: SRX026855 and SRX030853.

5.4.2 *De novo* assemblies of *O. tauri* genome

We used 3 *de novo* assemblers Velvet [30] (version 1.0.18), ABySS [29] (version 1.2.6) and CLCbio (version 4.06.beta) (`http://www.clcbio.com/products/clc-assembly-cell/`). These tools have a de Bruijn graph based algorithm and are well suited for short paired-end reads. During the scaffolding step, the number of paired-end reads required to join 2 contigs into a scaffold was set to 10 for both Velvet and ABySS. As there is no scaffolding step for CLCbio we used SSPACE [36]. Among the assemblies build with different k-mer sizes, the assembly with the highest weighted median length, N50, was kept for comparison between assemblers. The quality of *de novo* assemblies was assessed in terms of contiguity and correctness on scaffolds with a size greater than 500 bp. To remove bacterial sequences, contigs with less than 70% nucleotide identity (blastn) with available Mamiellales genome sequences were eliminated [388]. These comprised: *Bathycoccus prasinos*, *Micromonas pusilla*, *Micromonas* RCC299, *Ostreococcus* RCC809, *Ostreococcus lucimarinus* and *Ostreococcus tauri*. Contiguity statistics were the number of scaffolds, the N50, the assembly size and the percentage of the reference genome covered by the scaffolds (estimated from the number of aligned bases in the dnadiff report of NUCmer alignments see below).

5.4.3 Assessing assembly error rates of *de novo* scaffolds

Scaffolds were aligned against the reference genome using NUCmer from MUMmer v3.20 [421] with default options except for '-maxmatch -l 30 -banded -D 5'. A minimum exact-match anchor size was set to 30 bp and a minimum combined anchor length to 65 bp per cluster. Following Salzberg *et al.* [150], we discarded alignments with less than 95% identity, or more than 95% overlap with another alignment using delta-filter. From these alignments we tallied the correctness statistics using dnadiff [422] from MUMmer v3.20. The output was filtered

by removing all regions corresponding to repeated elements (TEs and tandem duplications). The correctness statistics are: the number of mis-joins (translocation, relocation or inversion) as defined in Salzberg *et al.* [150]. A mis-join is defined when subparts of a scaffold align on two different chromosomes (translocation), on the same chromosome in a different order (relocation) or are inverted compared to the reference (inversion). The error rate was computed as the mean number of mis-joins per scaffolds and as the proportion of scaffolds having at least one mis-join. To assess precisely how coding sequences were represented in *de novo* assemblies, we calculated the percent of aligned bases in the CDS from dnadiff after a NUCmer alignment of the scaffolds against the CDSs. The number of complete CDSs (start to stop) present in the assembly was obtained from the show-coords files (-l -c -b -T -o –r).

5.4.4 Improving a historical reference genome

Gap closing was performed in 4 steps using the OT-2009 dataset (1) *de novo* assembly, (2) IMAGE, (3) PCR sequencing, and (4) CRAC. *De novo* scaffolds recruitment to close gaps in the reference genome sequence was done as follows. *De novo* scaffolds were aligned onto the reference genome sequence using blastn. If the scaffold aligned onto the reference over 200 bp with 95% identity and with at least 50 bp on each side of a gap, the sequence of the scaffold was used to close the gap. As *de novo* assemblers may discard some informative low copy reads, we also used raw reads to improve the reference genome with two further steps. In a second step, we performed local iterative *de novo* assemblies using IMAGE [411] (version 2.1) and the 41 millions paired-end reads. We divided the genome into 597 super-contigs corresponding to n = 577 gaps and chromosomes (n = 20). IMAGE aligned the 41 M paired-end reads Illumina dataset against these super contigs using BWA (with default parameters). IMAGE subsequently gathered paired-end reads for which only one of the paired reads mapped at the end of one of two super contigs separated by a gap. If at least 10 paired-end reads were gathered, IMAGE performed a local assembly of these paired-end reads to elongate contigs iteratively.

Since the publication of the first version of the genome, primers have been designed manually to fill additional gaps, especially around coding regions. DNA from PCR were sent to sequencing platforms and this enabled 134 additional gaps in the updated genome version to be closed.

As a last step we used CRAC, a sensitive mapping method that uses a k-mer profiling approach of reads onto a reference genome [423]. CRAC first collects for each k-mer in the read its locations on the genome and its support (which is a proxy of the read coverage), then analyses both the variation of location and of support within the read: this enables the precise detection of deletions, insertions or translocations with DNA-seq data. This enabled us to extract paired reads that align on two different scaffolds and that could have been omitted in the previous approaches. We manually checked the positions mapped by these paired end reads on the reference genome and performed a manual assembly when possible. This enabled the filling of 34 additional gaps and the identification of two errors in the assembly that corresponded to inversions of one scaffold relative to its neighbouring scaffolds.

The mapping of the Illumina reads onto the reference (BWA, [419]) enabled the identification of nucleotide insertions/deletions (indels) variants compared to the reference genome sequence. A base in the reference was considered to be incorrect if at least a minimum of 10 reads scored the nucleotide differently (with both DNAseq and RNAseq). The incorrect nucleotide was then changed to the most occurring nucleotide if occurring in more than 90% of DNA reads. Previous analysis on SNP-calling on *O. tauri* mitochondrial and chloroplast genomes enabled us to estimate empirically that these coverage thresholds corresponded to 100% correct SNP predictions [402]. We applied the same cut-off for insertion/deletion correction of the reference genome sequence.

5.4.5 Genome evolution between 2001 and 2009

OT-2001 and OT-2009 reads were aligned on the reference genome with BWA with default parameters [419]. We used custom C scripts to scan the pileup files to call variants. There were 11 760 029 sites covered by a minimum of 10 reads and a maximum, which was chosen as 220 for OT-2001 and 256 for OT-2009 reads were retained for the analysis of the OT-2001 (corresponding to 125% of the average genome coverage for each

library), to discard low covered regions and possible duplicated regions in the reference genome. Candidate substitutions were identified when 99% of the OT-2001 reads were consistent with the reference nucleotide and 99% of the OT-2009 reads were consistent with the variant. This led to 12 candidate substitutions. In order to confirm each of these substitutions, we designed primers to sequence 100 bp each side of the substitution in the OT-2001 and the OT-2009 samples. The position of the substitution within the gene and the type of mutations (non-synonymous, synonymous, non-coding, nonsense) was obtained from manual inspection of the alignments of the 2001 and 2009 coding sequences. We used TMHMM to identify transmembrane domains [424, `http://www.cbs.dtu.dk/services/TMHMM-2.0/`]. The information about the gene families (number and presence of homologous genes) within sequenced green alga and land plants genomes, corresponding to the genes containing non-sense or frameshift mutations, were retrieved from the pico-PLAZA database [5]. The absence of a homologous gene was further confirmed by a tblastn against the genome sequence.

To investigate copy number variations between 2001 and 2009, we analysed the coverage over 50 bp windows along the chromosomes. Whenever we found a two fold or higher increase in coverage (as compared to the average genome coverage) with the OT-2009 reads, it was compared with the OT-2001 coverage.

5.4.6 Updated genome sequence annotation

RNAseq data was obtained from cells grown under diurnal LD cycles (12L12D). As most genes are expressed rhythmically in these conditions [425], we isolated RNA every 3 hours over a 24 hours cycle and pooled the samples for sequencing. RNA was extracted using the RNeasy-Plus Mini kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. Contaminating DNA was removed using RQ1 RNAse-free DNAse (Promega Corporation, Madison, US). Poly-A RNA was isolated and paired-end librairies were generated following the protocol from the Illumina mRNA-Seq Sample Prep Kit. Sequencing was carried out on a single lane of Illumina GIIx and 76 bp paired-end reads were obtained.

RNAseq data was used to guide the annotation procedure using the annotation pipeline developed at Gent University. Similar scripts can be downloaded from `https://mulcyber.toulouse.inra.fr/projects/eugene/`. The updated genomic sequence of *O. tauri* was annotated by using the EuGene [343, 426] gene finding system. Both Eugene (ab initio) as well as SpliceMachine [102] were specifically trained for *O. tauri* datasets. This pipeline integrates homology information derived from proteins sets of other microalgae from the Mamiellophyceae family; *Bathycoccus prasinos* RCC4222 (a clonal lineage re-isolated from RCC1105 [1]), *Micromonas pusilla* RCC299 and CCMP1545 [221], *Ostreococcus lucimarinus* [181], ESTs and full-length transcripts from *Ostreococcus tauri* that could be collected from NCBI, and all junctions present from mapping the present RNAseq dataset. Given the high density of the gene content in *O. tauri*, no RNAseq assembly was performed aiming at obtaining additional (full-length) transcripts. A trial-assembly of the RNAseq resulted in too many concatenated transcripts due to overlapping UTRs. A final thorough manually curation of the predicted gene models was performed by the authors using the ORCAE interface [110].

CHAPTER 6

GENOME RE-ENGINEERING FROM LAND TO SEA BY THE SEAGRASS *ZOSTERA MARINA*

*Jeanine L. Olsen, Pierre Rouzé, **Bram Verhelst**, Yao-Cheng Lin, Till Bayer, Jonas Collen, Emanuela Dattolo, Emanuele De Paoli, Simon Dittami, Florian Maumus, Gurvan Michel, Anna Kersting, Chiara Lauritano, Rolf Lohaus, Mats Töpel, Thierry Tonon, Kevin Vanneste, Mojgan Amirebrahimi, Janina Brakel, Christoffer Boström, Mansi Chovatia, Jane Grimwood, Jerry W. Jenkins, Alexander Jüterbock, Amy Mraz, Wytze T. Stam, Hope Tice, Erich Bornberg-Bauer, Pamela J. Green, Gareth A. Pearson, Gabriele Procaccini, Carlos M. Duarte, Jeremy Schmutz, Thorsten B. H. Reusch and Yves Van de Peer*

Abstract

Seagrasses colonized the sea [224] on at least three independent occasions to form one of the most productive and widespread coastal ecosystems on the planet [427]. The genome of *Zostera marina* (L.), the first marine angiosperm to be fully sequenced, reveals unique insights into the genomic losses and gains involved in achieving the structural and physiological adaptations required for its marine lifestyle, arguably the most severe habitat shift ever accomplished by flowering plants. As could be expected, a number of key angiosperm innovations were lost, such as the entire repertoire of stomatal genes [428], as well as key genes involved in the synthesis of terpenoids and ethylene signaling involved in aerial communication and resistance to insect herbivores through volatile organics. Additional reductions include the nucleotide binding site-leucine rich repeat (NBS-LRR) family involved in plant defense and loss of ultra-violet (UV) protection provided by UVR8 and phytochromes for far-red sensing. In contrast, seagrasses have also regained functions enabling them to adjust to full salinity and the altered light regimes of the marine environment. Their cell walls contain all of the polysaccharides typical of land plants but also polyanionic, low-methylated pectins and sulfated galactans, a feature shared with the cell walls of all macroalgae [429] and important for ion homoeostasis, as well as for nutrient uptake and O_2/CO_2 exchange through leaf epidermal cells. The co-existence of proton transporters and Na+/H+ antiporters maintain membrane potential against intrusion of Na+ from seawater and the pH imbalance created during carbonic anhydrase (CA)-mediated, bicarbonate transport for photosynthesis [430]. The *Z. marina* genome resource will significantly advance a wide range of functional ecological studies from adaptation of marine ecosystems under climate warming [431, 432] to unravelling the mechanisms of osmoregulation under high ambient salinities that may help in understanding the evolution of salt-tolerance in crop plants [433].

Contributions

- Bacterial contaminant filtering of JGI assembly
- Structural and functional genome annotation (including preliminary repeat detection)
- Manual gene curation
- Set-up and maintenance of the *Zostera*-section on the ORCAE platform
- Deposition of the *Z. marina* genome resource to NCBI GenBank
- Gene family clustering and comparative phylogenomics
- Collinearity and synteny comparisons
- Curation of chloroplast genome
- Figures *(Supplementary Figures 6.1, 6.2 and 6.7)*
- Tables *(Supplementary Table 6.3)*
- Writing manuscript segments (in respect to the topics mentioned above)

6.1 Introduction

Seagrasses are a polyphyletic assemblage of basal monocots belonging to four families in the Alismatales [224, 427]. Seagrasses have arisen independently at least three times within the Alismatales, a cosmopolitan and diverse, monophyletic clade of basal aquatic and marine monocots. The order comprises ~4500 species in 13 families (Angiosperm Phylogeny Group 2009) or 11 families (excluding Araceae and Tofieldiaceae) following Les & Tippery [434]. The seagrasses belong to four of five families depending on the taxonomic authority: the Hydrocharitaceae, Posidoniaceae, Cymodoceaceae/Ruppiaceae (separately or as one) and Zosteraceae. Hence, seagrasses constitute an ecological/functional grouping rather than a monophyletic clade as shown and reviewed in the molecular phylogenies of Les *et al.* [224] and Les & Tippery [434]. Fossil-calibrated, molecular clock estimates suggest that the three major clades of seagrasses (Hydrocharitaceae, Posidoniaceae plus Cymodoceaceae/Ruppiaceae, and Zosteraceae) evolved at roughly the same time between 40 -77 Mya.

Being an angiosperm with true root systems, seagrasses were able to occupy the previously empty niche in marine systems of shallow sedimentary shorelines. Seagrasses share many features with their freshwater sister taxa but they also possess uncommon and unique traits, which constitute evolutionary innovation in their return to the sea, an extremely rare set of events. As a functional group they provide the foundation of highly productive ecosystems present along the coasts of all continents except Antarctica, where they rival tropical rain forests and coral reefs in ecosystem services [435, 436]. In colonizing sedimentary shorelines of the world's ocean, seagrasses found a vast new habitat free of terrestrial competitors and insect pests but had to adapt to cope with new structural and physiological challenges related to full marine conditions.

Zostera marina (Zosteraceae) or eelgrass, is the most widespread species throughout the temperate northern hemisphere of the Pacific and Atlantic [229] *(Figure 6.1b)*, thereby providing an excellent model for functional ecological and evolutionary studies.

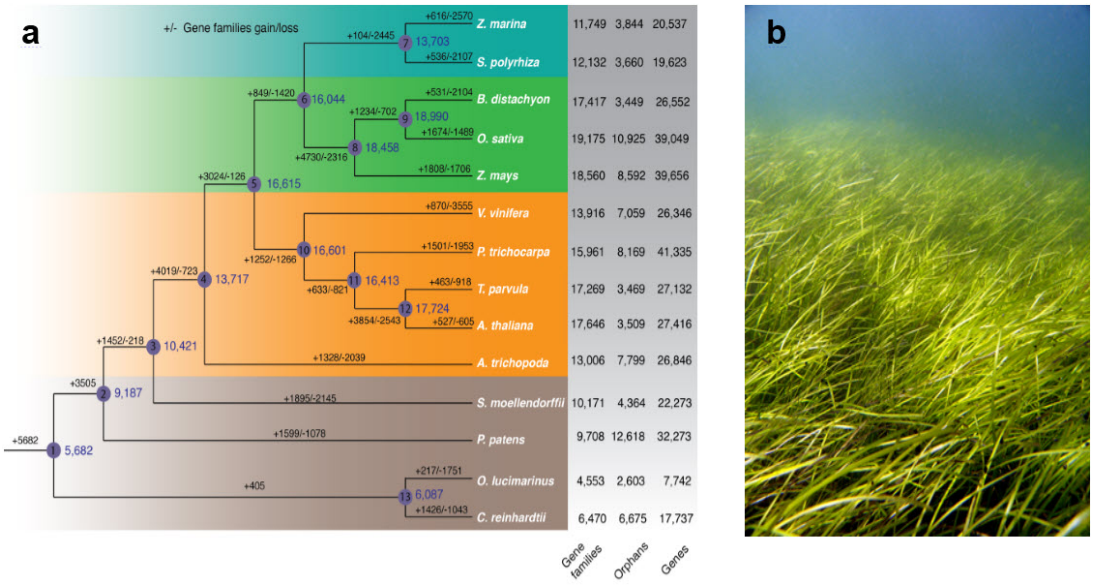


Figure 6.1. Phylogenetic tree and gene family expansion/contraction analysis for *Zostera marina* and 13 representatives of the Viridiplantae. (a) Gains and losses are indicated along branches and nodes. The number of gene families, orphans (single copy gene families) and number of predicted genes is indicated next to each species. Background shading (top to bottom) are Alismatales, other monocots, dicots, mosses/algae. **(b)** Typical *Zostera marina* meadow, Archipelago Sea, SW Finland (photo by C.B.).

6.2 Results and discussion

6.2.1 Sequencing and annotating the *Z. marina* genome

A clone of *Z. marina* was sequenced from the Archipelago Sea, SW Finland using a combination of fosmid-ends and whole-genome shotgun (WGS) approaches. In total, we obtained 0.14 Gb of Sanger sequence and 25.55 Gb of Illumina high quality reads representing ~47x genomic coverage *(Supplementary Table 6.1)*. Sequences were first assembled using ARACHNE2 [437], after which contigs were scaffolded using both mate-pair and fosmid–end libraries. Following removal of bacterial and plastid sequences, we obtained a final assembly of 202.3 Mb. Scaffold L50 length was 486 kb (N50:124) *(Supplementary Table 6.2)*. Complete sequencing of randomly selected fosmid clones indicated that the WGS assembly showed high base pair accuracy (<0.05% bp error). Further quality analysis indicated that 90% of the set of eukaryotic core genes (CEGMA) were present and 98% were partially represented, suggesting near completeness of the euchromatin component. We also assembled the chloroplast and partial mitochondrial genomes *(Supplementary Figure 6.5)*.

The 202.3 Mb *Z. marina* genome encodes 20,450 protein-coding genes, 86.6% (17,511 genes) *(Supplementary Table 6.3)* which is supported by transcriptome data from leaves, roots and flowers *(Supplementary Figure 6.1)*. The genes are located mostly in gene-dense islands separated by large stretches of repeat elements *(Supplementary Figure 6.2 and Supplementary Tables 6.3 and 6.4)*. Because the gene number and mean gene/exon/intron lengths of *Z. marina* are very similar to those of the recently sequenced duckweed, *Spirodela polyrhiza* (Alismatales, Araceae) [438], the main difference in assembly size can only be explained by a larger number of repeat elements between the gene islands *(Supplementary Figure 6.6)*. Almost 63% of the *Z. marina* non-gapped assembly consists of repeats *(Supplementary Table 6.3)*. Gypsy-type elements are predominant by contributing 32% of the repetitive elements, followed by Copia-type elements (20%). Sequence divergence analysis between copies and consensus sequences suggests that the genome retains copies from two distinct periods of invasion by Copia elements, but only one period for Gypsy elements *(Supplementary Figure 6.3a-c)*. The genes gained by *Z. marina* ('accessory') are located closer to transposable elements (TEs) and other genomic repeats on the genome sequence as compared to the conserved set of 'single copy' genes (Fisher's exact test, P <0.0001) indicating that TEs may have played a role in genic adaptation to the submarine environment. Remarkably, we found that proximal TEs (Gypsy-type elements) were more frequent in the gained genes as compared to conserved genes (Fisher's exact test, P < 0.0001), and that this subset of elements contains a high frequency of putative young copies.

6.2.2 MicroRNA analysis

Based on previously existing small RNA libraries and comparison with other sequenced plant genomes we identified 36 conserved micro-RNAs (miRNAs) with high confidence and their predicted targets. A novel variant of miR528 was found to be the only member of this miRNA family, demonstrating that this conserved miRNA is the only one that is ancestral to the entire monocotyledon lineage. The overall scenario emerging is that *Z. marina* did not take part in the subsequent birth of miRNAs that are common to several other monocots [439]; nor did it experience or retain traces of prominent miRNA duplications.

6.2.3 Whole-genome duplication

Analysis of K_S age distributions indicates that *Z. marin*a carries the remnants of an independent, ancient whole genome duplication (WGD) event *(Figure 6.2a)* [440]. Accordingly, ~9% of the *Z. marina* genome is found in duplicated segments, probably an underestimate due to the fragmented nature of the genome assembly *(Supplementary Figure 6.7)*. Analysis of K_S age distributions for *Z. marina* and *Spirodela polyrhiza* and comparison with the K_S distribution for orthologous genes of both species *(Supplementary Figure 6.8)* suggests that the WGD in *Z. marina* occurred independently from the double WGD reported for *S. polyrhiza* [438] and after both lineages diverged, somewhere between 135 and 107 Mya [441]. Phylogenomic dating [440] of the *Z. marina* WGD suggests that it occurred 72 - 64 Mya *(Figure 6.2b)*, thus at the time of initial diversification of a clade that includes three of the four families of seagrasses and in the timeframe of the Cretaceous-Paleogene

(K-Pg) boundary *(Figure 6.2c)*, which led to the extinction of 75% of all species and provided new ecological opportunities.

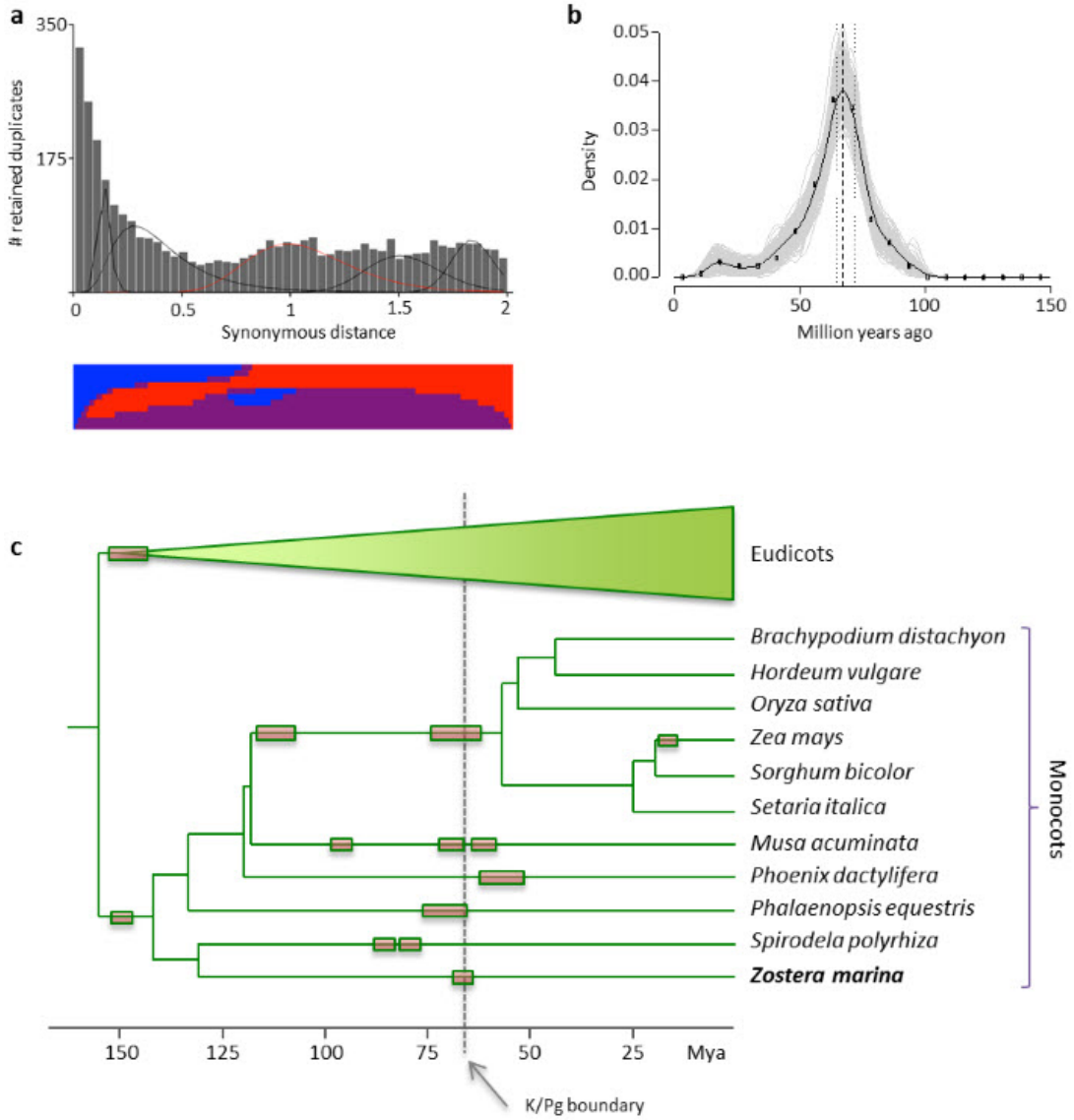


Figure 6.2. Ancient whole genome duplication (WGD). (a) K_S-based age distribution of the whole *Z. marina* paranome. The x-axis shows the synonymous distance until a K_S cut-off of 2 in bins of 0.04, containing the K_S values that were used for mixture modeling (excluding those with a $K_S \leq 0.1$). The component of the Gaussian mixture model plotted in red (as identified by EMMIX) corresponds to a significant WGD feature based on the SIZER analysis (other components are shown in black). The transition from the blue to the red at a K_S of ~1.00 in the SIZER panel indicates a significant change in the distribution and therefore provides evidence for an ancient WGD. **(b)**. Absolute age distribution obtained by phylogenomic dating of *Z. marina* paralogs. The solid black line represents the kernel density estimate (KDE) of the dated paralogs and the vertical dashed black line represents its peak, used as the consensus WGD age estimate, at 67 Mya. Gray lines represent the density estimates from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 64-72 Mya. The original raw distribution of dated paralogs is indicated by open circles. The y-axis represents the percentage of gene pairs. **(c)** Pruned phylogenetic tree with indication of WGD events (boxes) [442]. The Cretaceous/Paleogene (K/Pq) boundary is indicated by an arrow.

6.2.4 The seagrass adaptation to marine life

Clear signatures of loss and gain of gene families were mapped on a phylogenetic tree including *Z. marina* and 13 other species of the Viridiplantae *(Figure 6.1a)*. Losses and gains of Pfam domains were also mapped. Taken together, these losses and gains reflect the many unique aspects of seagrass adaptation to marine life. For example, all the genes involved in stomatal differentiation are absent from the *Z. marina* genome *(Figure 6.3a)*.

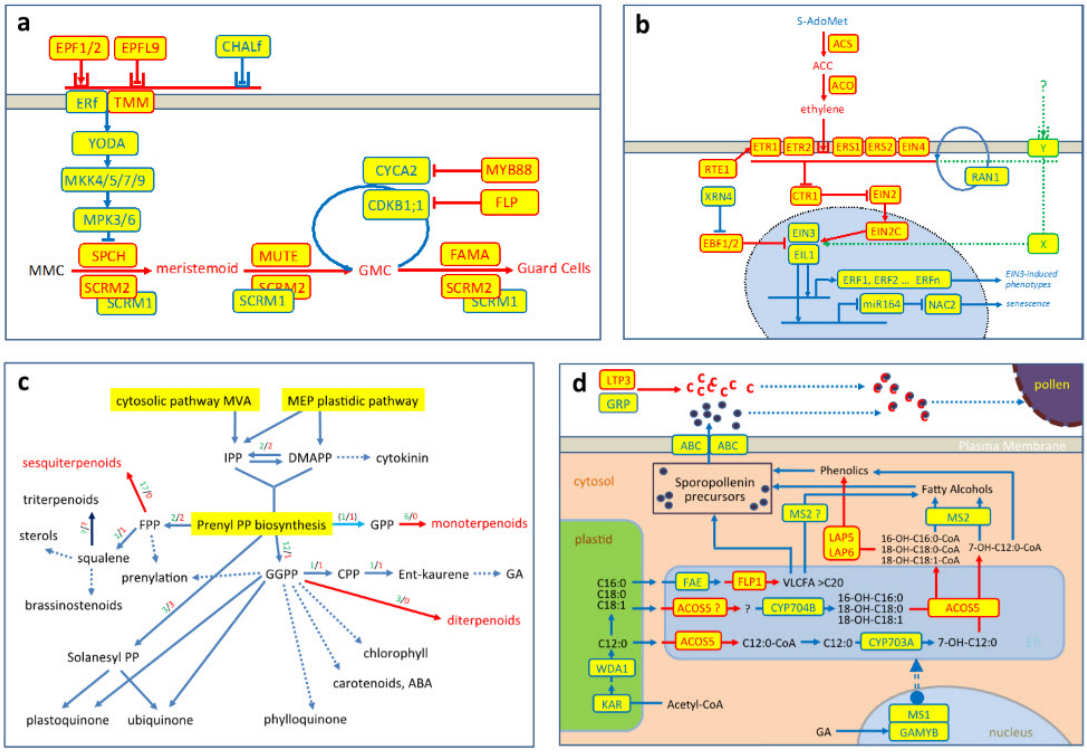


Figure 6.3. Reconstruction of pathways involved in the production of stomata, ethylene, terpene and pollen in *Z. marina*. (a) Stomata differentiation from Meristemoid Mother Cells (MMCs) to Guard Cells. Genes in red are missing. **(b)** Ethylene synthesis and signaling up to EIN2 have disappeared, whereas Ethylene Insensitive 3 (EIN3) and its downstream targets remain, suggestive of an alternative signaling pathway in seagrasses (green). Genes in red are missing. **(c)** Terpenoid biosynthesis showing that the pathways producing volatile terpenoids as secondary metabolites are absent, whereas the pathways essential for primary metabolism remain. **(d)** Sporopollenin biosynthesis genes, as well as processes and products that are absent from *Z. marina*, are shown in red; those still present in blue. Regulatory genes in the nucleus (grey oval) control downstream processes (arrows) in response to signalling coming from external stimuli through receptors on the plasma membrane (grey layer).

Ethylene signaling, terpenoids, stomata & defense genes

Genes comprising entire pathways encoding volatiles synthesis and sensing have also disappeared, such as those for ethylene [443] *(Figure 6.3b)*. However, the regulatory network downstream from Ethylene Insensitive 3 (EIN3) is still present, suggesting that the signaling has been rewired, presumably to a water diffusible molecule. Terpenoid genes are also drastically reduced to two *(Figure 6.3c)* as compared with 50 in *Oryza* and >100 in *Eucalyptus*, thus precluding synthesis of secondary volatile terpenes. Only aromatic acid decarboxylases (AAADs) genes were expanded. The loss of volatiles is also consistent with the loss of stomata, through which they are emitted for airborne communication and plant defense. The repertoire of defense-related genes such as NBS-LRR resistance genes is greatly reduced, which may be linked to a lower probability of infection of *Z. marina* due to the absence of stomata, which are a main entry point for pests and pathogen in terrestrial plants. Aside from their entry points, the unique repertoire of R-genes may indicate very different marine pathogens compared to land.

UV resistance

Land plants (Embryophyta) are often exposed to intense ultra-violet (UV) radiation and have developed light sensing protein receptors with protective and signaling functions. In contrast, *Z. marina* inhabits a light-attenuated, submarine environment where it must cope with shifted spectral composition, characterized by low penetration of UV-B, red, and far-red wavelengths [444]. Accordingly, *Z. marina* has lost ultra-violet resistance (UVR8) genes associated with sensing and responding to UV damage, as well as phytochromes associated with red/far-red receptors. Whereas photosystems (PSI and PSII) are similar to those of other plants including *S. polyrhiza*, members of the Light-harvesting complex B (LHCB) family are expanded in number, possibly in combination with non-photochemical quenching (NPQ), thereby enhancing performance at low light.

Osmoregulation

Seagrasses typically experience full-immersion salinities of 35 $g \times kg^{-1}$ [445], whereas land plants obtain water with usually low osmolarity via the rhizosphere. *Zostera marina* has adapted to this saline environment in several ways, playing on transporters and on unique features of its cell wall. The co-existence of proton transporters and Na+/H+ antiporters ensures maintenance of membrane potential against intrusion of Na+ from seawater and the pH imbalance created by carbonic anhydrase (CA)-mediated, bicarbonate transport for photosynthesis. Although *Z. marina* displays a typical repertoire of Na+ and K+ antiporters, one of six H+-ATPase (AHA) genes is strongly expressed in vegetative tissue and encodes a salt-tolerant H+-ATPase. Furthermore, *Z. marina* possesses three AHA genes (along with *Spirodela*) in a cluster unique to Alismatales, all three being expressed in male flowers, possibly regulating osmotic pressure in pollen during fertilization.

Algal-like cell wall & carbohydrate metabolism

*Zostera marin*a has re-evolved new combinations of structural and physiological traits related to the cell wall and its role in osmoregulation and ion homeostasis. With respect to the outer leaf coverings, synthesis of cutin-cuticular waxes to the outside of the leaf epidermis and suberin-lignin near the plasma membrane surround a cell wall matrix of (hemi)celluloses, low-methylated pectin (zosterin) and macroalgal-like sulfated polysaccharides [446]. The reduction in carbohydrate-related genes that modify the fine structure of cell wall hemicelluloses and pectins in *Z. marina* (716 genes compared to *Oryza* and *Arabidopsis* with >1100 genes) is not due to loss of pathways but rather to the large variation of these CAZyme families in plants. Available genomes of land plants (including the aquatic *Spirodela*) do not include carbohydrate sulfotransferases and sulfatases, suggesting that land plants have lost these genes as a key adaptation to terrestrial conditions [447, 448]. In contrast, *Z. marina* has regained the ability to produce sulfated polysaccharides with an expansion of aryl sulfotransferases (12 genes) homologous to aryl sulfotransferases from land plants. Sulfation facilitates water and ion retention in the cell wall to cope with desiccation and osmotic stress at low tide and, likewise, low methylation of zosterin correlates with the expanded pectin carbohydrate esterase 8 (CE8) family, increasing the polyanionic character of the cell wall matrix. We speculate that one or several aryl sulfotransferases have evolved because carbohydrate sulfatases have been shown to be active on artificial aryl compounds such as methylumbelliferyl-sulfate [449]. Osmotic equilibrium is further achieved in *Z. marina* by organic osmolytes (mainly sucrose, trehalose and proline) in combination with a small cytoplasm:vacuole volume ratio (10%) [430]. Up to 90% of fixed carbon is stored as sucrose, predominantly in the rhizome. Accordingly, sucrose synthase (SuSy) and sucrose transport (SUT) genes are expanded, as would be expected in “marine sugarcane”. Consistently, maltose- and starch-related genes are reduced although *Z. marina* still has the minimal number of genes to synthesize these storage carbohydrates.

Redox and stress-resistance genes

The repertoire of redox and other stress-resistance genes is typical for angiosperms with the exception of catalase (CAT), which is reduced to a single copy in *Z. marina*, and late embryogenesis abundant (LEA) and dehydrins are clearly under-represented in both *Zostera* and *Spirodela* relative to other genomes. In contrast, *Zostera* possesses an unusual complement of metallothioneins (MTs) compared to land plants. Aside from

their role as chelators, MTs may be involved in stress resistance; one of these (MT2L) is among the most highly expressed genes in *Z. marina (Supplementary Figure 6.4)*.

Exine-less pollen

Most freshwater alismatids (and also *Spirodela* [450]) possess pollen with an exine layer. Exine-less pollen [451] is characteristic of *Z. marina* and all other seagrasses except *Enhalus acoroides*, which is surface pollinated and appears to be a convergent adaptation in the transition from aerial to submerged pollination systems. Ten genes specifically involved in biosynthesis and modification of the pollen exine coat are missing; all other genes involved in the development of viable pollen remain intact *(Figure 6.3d)*. *Z. marina* and most of the other seagrasses have also evolved a unique filiform pollen that winds around the bifurcate stigmas in a purely abiotic pollination process [452]. Sexual reproduction of completely submerged male and female flowers is entirely hydrophilic and no longer mediated by wind or insects. Finally, MADS-box gene transcription factors are also reduced to 50 in *Z. marina*, which is most likely related to the highly reduced flowers (also a feature in *S. polyrhiza*) that lacks the first two whorls of specialized floral leaves, calyx and corolla.

6.3 Conclusions

So far, genomic efforts in monocots have mainly targeted agriculturally important species and those of interest for biofuel feed stocks and re-mediation; most are true grasses and, until recently, all have been terrestrial or freshwater. As a bridge between the terrestrial and marine domains, the *Z. marina* genome provides a major resource for understanding how structural and physiological functions have been lost, gained or re-engineered in adapting to the submerged life-style of high salinity, low and spectrally-shifted light, the exchange of gasses and nutrient uptake through leaves rather than stomata, and a unique underwater fertilization *(Figure 6.4)*.

An increasing proportion of the world population inhabits the coastal zone. This impinges multiple pressures on ecosystems including seagrass beds [453, 454], which in turn compromises the ecosystem services they may provide, including provisioning of harvest-able fish and invertebrates, nutrient retention and erosion control. Elucidating the complex adaptations of the seagrass *Z. marina* to ocean waters will further advance our understanding of the evolution of salinity tolerance that may inform assisted breeding of terrestrial crop plants [433]. The first description of the full genome sequence of a plant that has successfully colonized the sea provides multiple opportunities to study the genomics of plant adaptation to climate change [431, 432], and for developing molecular indicators of their physiological status [455] in the context of seagrass conservation, as these unique ecosystems rank, unfortunately, among the most threatened on Earth [453, 454].

6.4 Materials and methods

6.4.1 Plant material and DNA preparation

A single genotype/clone of *Zostera marina* (referred to as the 'Finnish clone') was harvested on 26 August 2010 at 2 m depth at Fårö Island (lat. 59° 55.234' long. 21° 47.766') located in the northern Baltic Sea, Finland. Plant material was transported to the lab in seawater, cleaned and further processed. Care was taken to use leaf-meristem tissue harvested from the inner layer of basal shoots to minimize bacterial/diatom contamination. Tissues were immediately frozen in LN2 and stored at -80°C for later DNA and RNA extraction. Monoclonality was verified by genotyping 40 ramets of the mega-clone with six highly polymorphic, microsatellite loci [456]. There was no evidence for polyploidy [451, 457, 458] (*Z. marina* is 2n=12) or somatic mutations [459] as assessed by multiple peaks in the microsatellite chromatograms. Tissue was subsequently sent on dry ice to Amplicon Express (Pullman, WA, USA) for HMW DNA extraction using a CTAB isolation method modified by R.Meilan (unpublished) but available from him (rmeilan@purdue.edu), based on the original method [460]. Following QC according to JGI guidelines, the DNA was shipped to JGI for library and sequencing preparation.

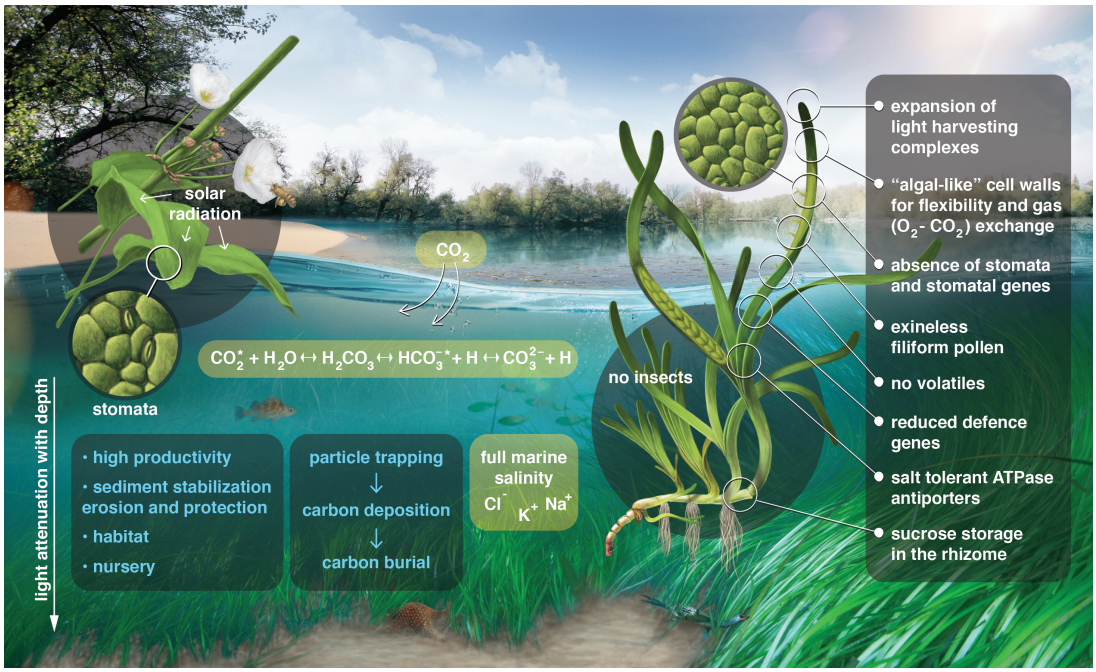


Figure 6.4. Conceptual summary of physiological and structural adaptations made through re-engineering of the genome by *Z. marina* in its return to the sea. Physical conditions include full salinities of 35 g*kg^{-1}, requiring changes in osmoregulation; and diminished, as well as spectrally shifted light along the water column that requires altered light harvesting mechanisms. Morphological adaptations include stomata-less leaves and 'algal-like' cell walls that function in gas (O_2-CO_2) and nutrient exchange, as well as osmoregulation and ion homeostasis. Plant defense genes needed for terrestrial pests and pathogens are reduced, as are volatiles. Submarine fertilization led to convergent evolution of exineless and filiform pollen. Ecosystem services include particle trapping and carbon burial, as well as coastal erosion protection, extremely high productivity and nursery functions.

6.4.2 Genome sequencing and assembly

One fosmid library was generated for end sequencing. The fosmid reads were sequenced with standard Sanger sequencing protocols at the HudsonAlpha Institute in Huntsville, Alabama, USA for a total of 0.29x coverage. Illumina reads were sequenced by the Illumina MiSeq/HiSeq machines at the Department of Energy's Joint Genome Institute (JGI), Walnut Creek, California, USA using standard protocols. Two Illumina fragment libraries (6.62 Gb), one 2kb JGI mate pair library (3.57 Gb), one 4kb JGI mate pair library (3.41 Gb), two 8kb JGI mate pair libraries (11.94 Gb) were sequenced on Illumina. One 35kb fosmid library was sequenced on both ends with Sanger sequencing for a total of 194,303 Sanger reads. A total of 25.55 Gb of Illumina and 0.14 Gb of Sanger sequence was obtained. Prior to assembly, all reads were screened against mitochondria, chloroplast, and Illumina controls. Reads composed of >95% simple sequence repeats were removed. For the Illumina reads, reads in 2x250 libraries <75bp were discarded, and reads in 2x150 libraries <50bp were discarded after trimming for adapter and quality (q<20). An additional deduplication step was performed on the mate pairs that identifies and retains only one copy of each PCR duplicate. A total of 212,101,273 reads *(Supplementary Table 6.1)* were assembled using our modified version of ARACHNE v.20071016 [437]. Subsequent directed ARACHNE modules were applied to collapse adjacent heterozygous contigs. The entire assembly was then run through another ARACHNE process starting at STAGE 6 REBUILDER. This produced 15,747 scaffold sequences (30,723 contigs), with a scaffold L50 of 409.5 kb, 613 scaffolds larger than 100 kb, and total genome size of 237.5 Mb.

Scaffolds were screened against bacterial proteins, organelle sequences, GenBank NR (nr_prot) and RefSeq protein databases, and removed if found to be a contaminant. Scaffolds consisting of prokaryotes, chloroplast, mitochondria and unanchored rDNA were removed. Additionally, short (<1kb) scaffolds or scaffolds containing highly repetitive sequence (>95% 24 mers found more than four times in large scaffolds) or alternative

haplotypes were removed as well. Furthermore, after repeat analysis and gene prediction, all scaffolds were subjected to a filtering process (based on NCBI nr_prot + NCBI taxonomy database) to eliminate remaining bacterial (and other) contaminants *(Supplementary Table 6.2)*.

Assembly validation was performed using a set of 12 fully sequenced fosmid clones. In four of the 12 fosmid clones, full length alignments were not found due to fragmentation in the region of the fosmid clone. In five of the remaining eight fosmid clones, the alignments were of high quality (< 0.50% bp error). The overall base pair error rate (including marked gap bases) in the fosmid clones that aligned to full length was 0.28% (714 discrepant bp out of 253,332 bp). Note that two fosmid clones (16248, 16249) contributed nearly 81% of the discrepant bases. This probably occurred in polymorphic regions of the genome where the haplotype in the fosmid did not match the haplotype in the reference. There are several indels of various sizes in the clone and assembly, typical of a region of degraded transposons.

6.4.3 Annotation of repetitive sequences

Two complementary approaches were used to identify repetitive DNA sequences in the *Z. marina* genome. With respect to masking repeats prior to gene prediction analysis, a *de novo* repeat identification was carried out with REPEATMODELER (version open-1.0.7; [79, `http://www.RepeatMasker.org`]) to identify repeat boundaries and build consensus models from which potential over represented, non-transposable element, protein coding genes were removed. REPEATMASKER (ver. open-4.0.0, WUBLAST) was used in combination with this custom repeat library to mask the assembly and prepare it for gene prediction with EuGene.

Furthermore, in order to perform a qualitative and quantitative analysis of repeats with greater resolution [461] the genome assembly was processed for *de novo* repeat detection using the TEdenovo pipeline from the REPET package (v2.2 [80]; parameters were set to consider repeats with at least five copies). The consensus sequences generated by TEdenovo were then used as probes for whole genome annotation by the TEannot [462] pipeline from the REPET package v2.2. The consensus repeat sequences were classified using PASTEC [81]. Comparing the genomic positions of transposable elements (TEs) to those of exons from the set of predicted genes enabled to identify that 909 gene predictions most likely represent TEs and these were filtered from the gene set. The REPET package v2.2 was also used to annotate repetitive elements in the *Spirodela polyrhiza* genome assembly with the same parameters as for *Z. marina*.

6.4.4 Transcriptome library preparation, sequencing and assembly

Leaf, root and flower tissues were separately frozen in liquid nitrogen immediately following harvest from either ambient (field collected) or experimental (mesocosm) conditions. Overall, we obtained between nine and 20 million high quality reads from each of the flower-leaf-root replicate libraries; and for the Finnish clone library, 148.5 million high quality reads were retrieved.

The *de novo* assembly protocol was adapted from [463]. We pooled replicates of each tissue together except for the two leaf tissue libraries, which were kept separate and performed *de novo* transcriptome assembly for each tissue using TRINITY (ver. 2014-07-17) [463] with digital normalization option ON to normalize input read coverage. Frame shift errors and insertion/deletion errors in the assembled transcripts were corrected by FRAMEDP [464]. Because a *de novo* assembly still generates many spurious transcripts, we used the transcript expression value to remove low quality contigs. We used the RSEM pipeline [465] to obtain the contig expression values and removed contigs with Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) value < 1 and IsoPct (percentage of expression for a given transcript compared with all expression from that TRINITY component) < 1. In total, we obtained between 39K and 53K assembled contigs from each library, and 52K contigs from the Finnish clone library. Prior to mapping the genome sequence and the predicted genes, we used CD-HIT (ver. 4.6.1) [466] to collapse redundant contigs, which resulted in 79,134 low redundant transcript contigs.

6.4.5 Differential gene expression analysis

High quality RNAseq reads were mapped to the genome assembly v2.1 by TOPHAT [467]. Differential gene expression analysis was performed by the CUFFLINKS pipeline [467] based on the *Z. marina* v2.1 gene models by converting the number of aligned reads into FPKM values. Genes with significant expression difference (log2 >2) were selected for further investigation by GOSTATS [468] to perform Gene Ontology (GO) term enrichment analysis with $p \leq 0.05$.

6.4.6 MicroRNA analysis

Genomic precursors of known miRNAs were mapped on the *Z. marina* genome following the procedure described in Zhang *et al.* [469] for the *maize* genome. miRNA entries from the miRBase database (release 21, 2014) were aligned to the chromosomes of the *Z. marina* genome. Up to three mismatches were allowed in the alignment, using SEQMAP [470]. In parallel, novel potential DCL1/AGO1-dependent miRNAs were enriched by selecting 5'-U 20-22 nt small RNAs from three different sequenced libraries from *Z. marina* described in Chavez Montes *et al.* [439]. A subset of these small RNAs with abundance ≥ 10 TPM (Transcripts Per Million) was retained and aligned to the genome with no mismatches. From every locus, we extracted two ~200-nt regions surrounding each aligned miRNA or candidate (from -30 to +160 and from -160 to +30 nucleotides relative to the putative miRNA start or end coordinate, respectively). Minimum energy RNA secondary structures were predicted for each region using the RNAFOLD program of the VIENNA RNA 1.8.5 package (`http://www.tbi.univie.ac.at/~ivo/RNA/`) using default settings.

In addition, small RNAs from the three sequenced libraries were mapped on these regions, allowing no mismatches, in order to pre-select putative miRNA loci that showed evidence of expression in the three plant tissues analysed. We evaluated RNA structure and small RNA alignment in all the regions based on: 1) dominance of plus-stranded small RNAs; 2) position of the most abundant small RNAs relative to the predicted miRNA coordinates; 3) prevalence of 20-22 nt small RNAs in the predicted miRNA locus; 4) position of the putative miRNA with the stem-loop structure; and 5) absence of oversize (≥ 3 nt) bulges in the miRNA/miRNA* alignment. After reduction of overlapping loci to a non-redundant set and removal of stem-loop structures with the wrong orientation compared to miRNAs registered in miRBase, we manually inspected the remaining loci to further evaluate them according to the miRNA annotation criteria proposed by Meyers *et al.* [471]. Stringency was relaxed when small RNA expression data strongly indicated the presence of miRNA loci that did not meet the whole set of criteria. Novel miRNA precursors overlapping with TEs or other repetitive elements were filtered out.

Potential miRNA targets were identified in silico using the generic small rna-transcriptome aligner GSTAR from the CleaveLand package (version 4) [472]. Predicted targets were accepted with a Allen score < 4 or a MFE (Minimum Free Energy) ratio ≥ 7.5.

6.4.7 Gene prediction

Training of the gene prediction programs started with the collection of high quality EST information from the *Dr. Zompo* database [473, `http://drzompo.uni-muenster.de/`]. EST information was used, for example, to train the splice predictor SPLICEMACHINE [102]. Detection of conserved splice sites was further investigated by RNAseq splice junctions (count > 10) to construct a WAM model in EuGene (ver.4.1) [101]. Coding-potential was modeled with an Interpolated Markov Model (IMM) constructed from the BLASTX alignments of proteins from the PLAZA v2.5 database [474]. An additional protein 'monocot' Markov Model was built based on the protein sequences from *Brachypodium*, *maize* and *sorghum*. Starting from EST and protein alignments, a set of 215 gene models was manually constructed and curated using the genome browser GENOMEVIEW [108]. The 215 models were then used as a training set for EuGene in order to optimize the different splice site and coding-potential models, as well as the weights for the extrinsic EST and homology evidence. An overall fitness score of 80.1% was achieved, which is high enough to obtain reliable results without overfitting. GENEMARK [475] and AUGUSTUS [476] were separately trained (using the same input data as EuGene) and their predictions were integrated with EuGene using a custom script to evaluate the best gene structure at each locus. All gene

models were automatically screened to highlight possible erroneous structures (e.g., in-frame stop codons, deviating splice junctions, etc.) and manually curated.

Transfer-RNA gene models were predicted by TRNASCAN-SE (v1.31) [113] and their structures were verified with INFERNAL (v1.1rc1, rfam11 covariant model database) [112]. For each gene, UTRs were assigned by identifying a set of ESTs and RNAseq assemblies that uniquely overlapped with it. We subsequently selected the longest mapped transcript on either end of the predicted coding sequence and designated the section outside the coding sequence as the UTR. Finally, all genes were uploaded to the ORCAE platform [110, `http://bioinformatics.psb.ugent.be/orcae`], enabling all members of the consortium to refine and curate the gene model and assign gene function.

A list of protein domains, as well as the derived Gene Ontology (GO) terms and KEGG pathway identifiers were generated using an INTERPROSCAN (ver.5.2.45) [477] analysis and are available in ORCAE. More specifically, gene functional descriptions were added either manually by consortium expert scientists or automatically through sequence homology searches. The automated method relies on either the Enzyme Commission (EC) number reported by INTERPROSCAN to retrieve the enzyme name (if available), or an homology-based search where the functional description of the homolog is transferred if it meets specific criteria. After a BLASTP search against UNIPROTKB/SWISS-PROT [478] we filter out hits that are below 60% identity and 70% query/hit coverage. Although such high stringency on percent identity and sequence coverage reduced the available number of functional descriptions, it will reduce the false positive prediction rate, as desired here.

6.4.8 Construction of age distributions and WGD analyses

K_S-based age distributions were constructed as previously described in Vanneste *et al.* [479]. Briefly, the K_S values between genes were obtained through maximum likelihood estimation using the CODEML program [480] of the PAML package (v4.4c) [481]. Gene families for which K_S estimates between members did not exceed a value of 5 were subdivided into subfamilies. For each duplicated gene in the resulting phylogenetic gene tree, obtained by PHYML [358], all m K_S estimates between the two child clades were added to the K_S distribution with a weight 1/m, so that the weights of all K_S estimates for a single duplication event summed to one. Mixture modeling was used to confirm a WGD signature in the K_S distribution *(Figure 6.2 and Supplementary Figure 6.7)*, for which all duplicates with K_S values ≤ 0.1 were excluded to avoid the incorporation of allelic and/or splice variants, while all duplicates with K_S values > 2.0 were removed because K_S saturation and stochasticity can mislead mixture modeling above this range [479].

Absolute dating of the identified WGD event was performed as described previously [440, 442]. Briefly, paralogous gene pairs located in duplicated segments (anchors) and duplicated pairs lying under the WGD peak (peak-based duplicates) were collected for phylogenetic dating. Anchors, assumed to be corresponding to the most recent WGD, were detected using I-ADHORE 3.0 [346, 482]. Only a low number of duplicated segments and hence anchors could be identified, most likely because of the fragmented assembly of *Z. marina*. However, the identified anchors did confirm the presence of a broad WGD peak between a K_S of 0.8 and 1.6 (data not shown). For each WGD paralogous pair, an orthogroup was created that included the two paralogs plus several orthologs from other plant species as identified by INPARANOID (v4.1) [483] using a broad taxonomic sampling: one representative ortholog from the order Cucurbitales, two from the Rosales, two from the Fabales, two from the Malpighiales, two from the Brassicales, one from the Malvales, one from the Solanales, two from the Poales, one ortholog from *Musa acuminata* [484] (Zingiberales), and one ortholog from *Spirodela polyrhiza* [438] (Alismatales). In total, about 180 orthogroups from anchor pair duplicates and peak-based duplicates were collected. The node joining the two *Z. marina* WGD paralogs was then dated using the BEAST v1.7 package [485] under an uncorrelated relaxed clock model and a LG+G (four rate categories) evolutionary model. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the consensus APGIII phylogeny [486]. Fossil calibrations were implemented using log-normal calibration priors on the following nodes: the node uniting the Malvidae based on the fossil *Dressiantha bicarpellata* [487] with prior offset=82.8, mean=3.8528, and SD=0.5) [488], the node uniting the Fabidae based on the fossil *Paleoclusia chevalieri* [489] with prior offset=82.8, mean=3.9314, and SD=0.5 [490], the node uniting the Alismatales (including *Z. marina* and *Spirodela polyrhiza*) with the other monocots based on the oldest fossil monocot pollen, *Liliacidites* [491, 492] from the Trent's Reach locality (Virginia, USA), with prior offset=125, mean=2.0418, and SD=0.5 [441, 493]

and the root with prior offset=124, mean=4.0786, and SD=0.5 [494]. The offsets of these calibrations represent hard minimum boundaries, while their means represent locations for their respective peak mass probabilities in accordance with some of the most recent and taxonomically complete dating studies available for these specific clades [441, 495]. A run without data was performed to ensure proper placement of the marginal calibration prior distributions [496]. The Markov Chain Monte Carlo (MCMC) for each orthogroup was run for 106 generations, sampling every 1,000 generations resulting in a sample size of 104. The resulting trace files of all orthogroups were evaluated manually using TRACER v1.572 with a burn-in of 1,000 samples to ensure proper convergence (minimum ESS for all statistics at least 200). In total, 169 orthogroups were accepted and all age estimates for the node uniting the WGD paralogous pairs were then grouped into one absolute age distribution *(Figure 6.2)* – too few anchors were available to evaluate them separately from the peak-based duplicates) – for which kernel density estimate (KDE) and a bootstrapping procedure were used to find the peak consensus WGD age estimate and its 90% confidence interval boundaries, respectively.

Intra- and inter-genomic collinearity was investigated using MCScanX [497] based on a BLASTP search of all genomic protein coding genes with an E-value cutoff of e-10. Only one large duplicated segment was detected, which was most likely due to the fragmented assembly of *Z. marina*; only 27 scaffolds had a size larger than 1 Mb, accounting for only 23.4% of all protein coding genes. We therefore additionally used I-ADHORE (v3.0) [346] to investigate genomic collinearity by including all possible scaffolds.

6.4.9 Gene family comparisons

Protein sets were collected for 14 species: *Z. marina* (ORCAE v2.1), *Arabidopsis thaliana* (TAIR10), *Thellungiella parvula* (http://thelungiella.org), *Populus trichocarpa* (Phytozome v9.0), *Vitis vinifera* (Phytozome v9.0), *Amborella trichopoda* (http://amborella.huck.psu.edu), *Oryza sativa* japonica (Phytozome v9.0), *Zea mays* (Phytozome v9.0), *Brachypodium distachyon* (Phytozome v9.0), *Spirodela polyrhiza* (http://mocklerlab.org), *Selaginella moellendorffii* (Phytozome v9.0), *Physcomitrella patens* (Phytozome v9.0), *Chlamydomonas reinhardtii* (Phytozome v9.0), and *Ostreococcus lucimarinus* (ORCAE v6/3/2013). These species were selected in order to provide a phylogenetic representation traversing green algae, basal plants, monocots, and dicots. Following an 'all-vs-all' TimeLogic Decypher TERA-BLASTP (Active Motif Inc., Carlsbad, CA; e-value threshold 1e-3, max hits 500) comparison, ORTHOMCL (v2.0; mcl inflation factor 3.0) [498] was used to delineate gene families. Confidence in establishing gene losses in Zostera was enhanced by using a combination of reciprocal blast, TBLASTN, reannotation of *Spirodela* (and other monocot genes), and careful phylogenetic analysis. ORTHOMCL results and related protein resources are available in the ORCAE download section.

To further understand gene family expansion or contraction in *Z. marina* in comparison with other sequenced genomes, gene family sizes were calculated for all gene families (excluding orphans and species-specific families). The number of genes per species for each family was transformed into a matrix of z-scores in order to center and normalize the data. The first 100 families with the largest gene family size in *Z. marina* were selected. The z-score profile was hierarchically clustered (complete linkage clustering) using Pearson correlation as a distance measure. The functional annotation of each family was predicted based on sequence similarity to entries in the InterProScan and Pfam protein domain database where more than 30% of proteins in the family share the same protein domain. The phylogenetic profile and phylogenetic tree topology provided at PLAZA [499] were used to reconstruct the most parsimonious series of gene gain and loss events. The DOLLOP program from the PHYLIP package [500] was used to determine the minimum gene set at ancestral nodes of the phylogenetic tree. The DOLLOP program is based on the Dollo parsimony principle, which assumes that novel gene(s) families arise exactly once during evolution but can be lost independently in different phylogenetic lineages.

6.4.10 Search for presence/absence of orthologs for specific genes and families

A dedicated search for orthologs/homologs was performed for genes and proteins involved in stomata differentiation, volatile biosynthesis and sensing with focus on ethylene and terpenes, as well as genes involved in male flower specification and pollen differentiation. To this end, queries were chosen from documented genes involved in these pathways (usually from *Arabidopsis* but occasionally from *Oryza*, *Zea* and *tomato*). Next, the

search for homologs in *Zostera marina*, *Spirodela polyrhiza*, *Oryza sativa* japonica and *Arabidopsis thaliana* (when not used as a query) was performed using BLASTP. To avoid missing or poorly annotated genes a TBLASTN search was conducted with the queries against the *Zostera marina* and *Spirodela polyrhiza* genomes. Putative orthologs were identified based on reciprocal BLASTP searches with *Arabidopsis* (or the other queries). Due to species-specific duplications, this sometimes produced a number of paralogous genes orthologous to the query, or vice versa. To further confirm correct orthology assignments, phylogenetic trees were built using a broader sampling of protein sequences from both the query species and the three target species. Ambiguously aligned sequences (especially due to indels) were checked manually and corrected or removed.

6.5 Supplementary Information

This section contains selected segments of supplementary figures and tables most relevant to this chapter and the topic of this dissertation.

Library	Sequencing Platform	Average Insert Size	Read Number	Assembled Sequence Coverage
MONE	Illumina fragment	763 ± 21	17,376,230	10.32
MTWO	Illumina fragment	775 ± 22	17,321,160	8.80
IWHB	Illumina mate-pair	1,820 ± 190	37,108,670	8.44
IUSG	Illumina mate-pair	3,520 ± 436	36,295,144	7.81
NUUS	Illumina mate-pair	7,678 ± 828	45,575,590	5.20
NUUP	Illumina mate-pair	8,133 ± 999	58,230,176	6.84
XTP	Sanger	35,450 ± 4,655	194,303	0.29
Total		NA	212,101,273	47.70

Supplementary Table 6.1. Genomic libraries included in the *Zostera marina* genome assembly and their respective assembled sequence coverage levels in the final release version 2.1.

V2.1 assembly	Scaffolds	Contigs
Sequences	2,228	12,583
Total Length	203,914,448	191,659,986
Max. Length	2,654,544	642,312
Min. Length	1,000	200
N50	124	623
L50 (b)	485,578	79,958

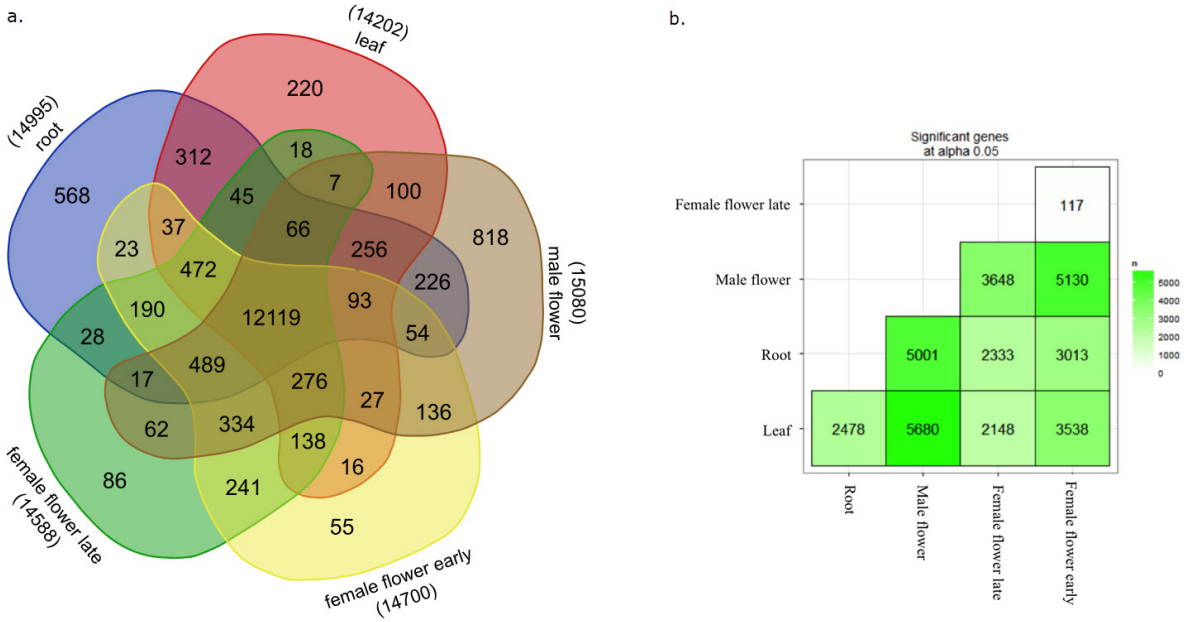
Supplementary Table 6.2. Summary of assembly statistics for *Z. marina* assembly V2.1.

	A. thaliana	*O. sativa*	*B. distachyon*	*S. polyrhiza*	*Z. marina*
Genome assembly					
Genome size (Mb)	119	374	271	128	203
Assembly status	5 chr	12 chr	N50: 3 L50: 59.3 Mb	22 pseudochr	N50: 124 L50: 496 Kb
Genome annotation[1]					
No. protein-coding genes	27,416	39,049	26,522	19,623	20,450
No. multi-exon genes (%)	20,607 (75.2%)	28,071 (71.9%)	20,317 (76.5%)	15,817 (80.6%)	15,985 (78.2%)
No. single-exon genes	6,809 (24.8%)	10,978 (28.1%)	6,235 (23.5%)	3,806 (19.4%)	4,465 (21.8%)
Avg. gene density (kb/gene)	4.4	9.6	10.2	6.5	9.9
Avg. gene / CDS length (bp)	1,867/1,218	2,329/1,064	2,851/1,284	3,458/1,108	3,301/1,177
Avg. exon / intron length (bp)	237/152	258/401	256/384	213/559	227/443
Avg. exons per gene	5.1	4.1	5.0	5.2	5.2
Avg. Intergenic (bp)	2,211	16,382	17,177	3,926	5,029
transposable elements					
Overall TE content (%)	24[2]	35[3]	21.4[4]	13[5]	63
Class 1 LTR Gypsy/Copia (%)	5.9/1.6	10.9/3.9	4.8/16.1	6.1/1.7	32/20
Class 1 LINE/unknown (%)	1/0.1	1.1/3.4	1.9/0.5	-/5.3	7/1
Class 2 DNA transposon (%)	12.1	12.9	4.8	-	6
Unclassified (%)	0.2	1.8	-	-	11

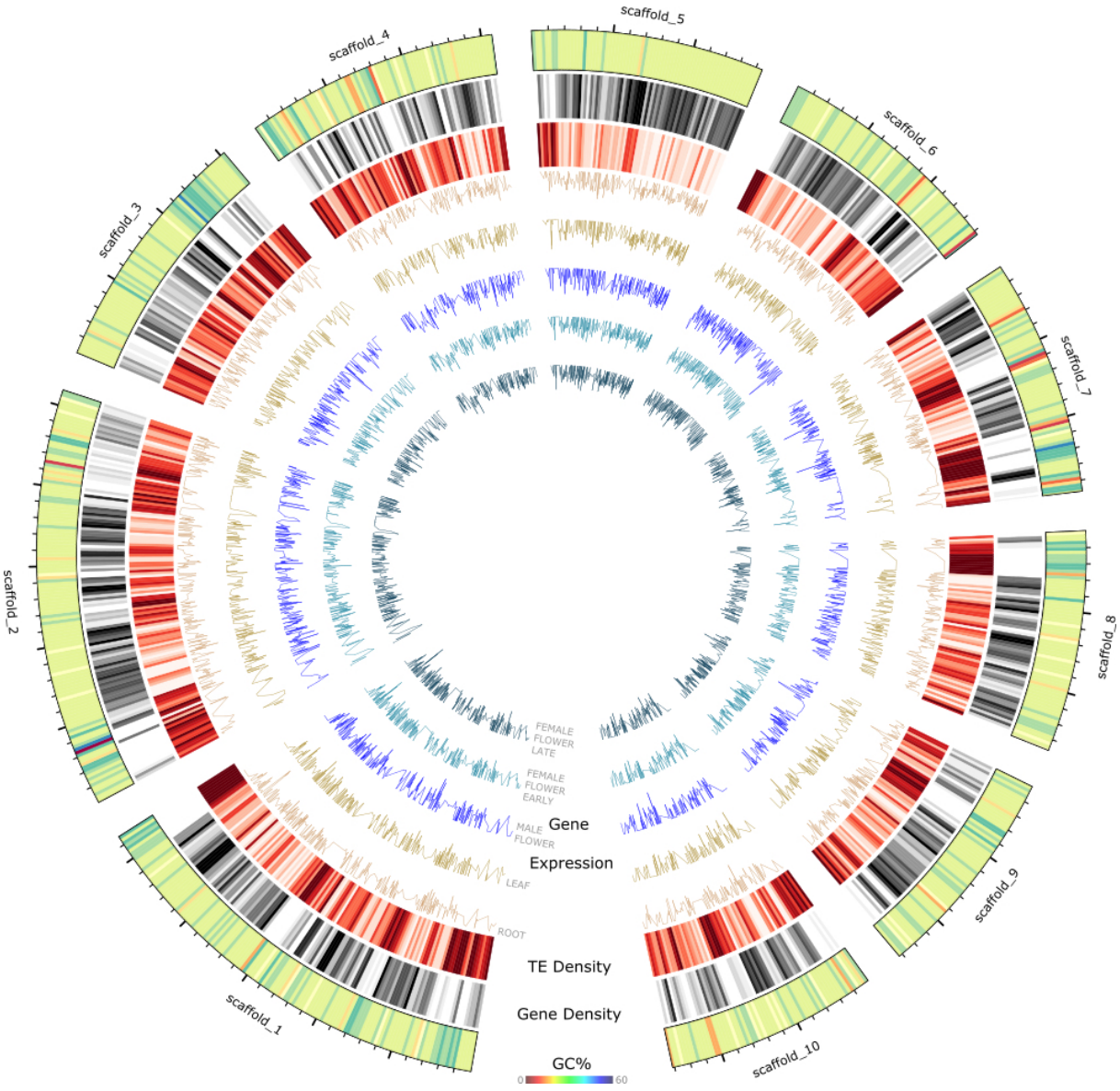
Supplementary Table 6.3. Summary of genes and transposable elements in *Z. marina* and other plant genomes. Sources: [1]Phytozome 9 annotation for *A.thaliana*, *O. sativa*, *B. distachyon* and *S. polyrhiza* gene models. [2][501]. [3][502]. [4][503]. [5][438].

Type	Total bases
Copia	23,632,031
Gypsy	39,114,803
LINE	8,670,051
SINE	21,222
putative retrotransposon	1,613,655
DNA transposons	6,857,309
Unclassified	12,827,267
Host Gene	15,257,469
satellite repeats	12,406,906

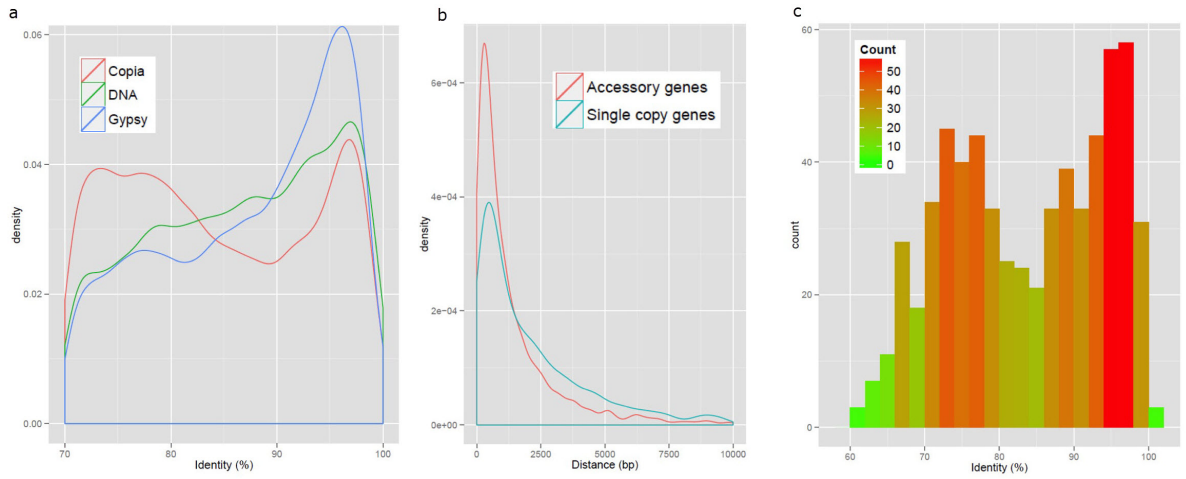
Supplementary Table 6.4. Summary of transposable elements annotation in *Z. marina*.



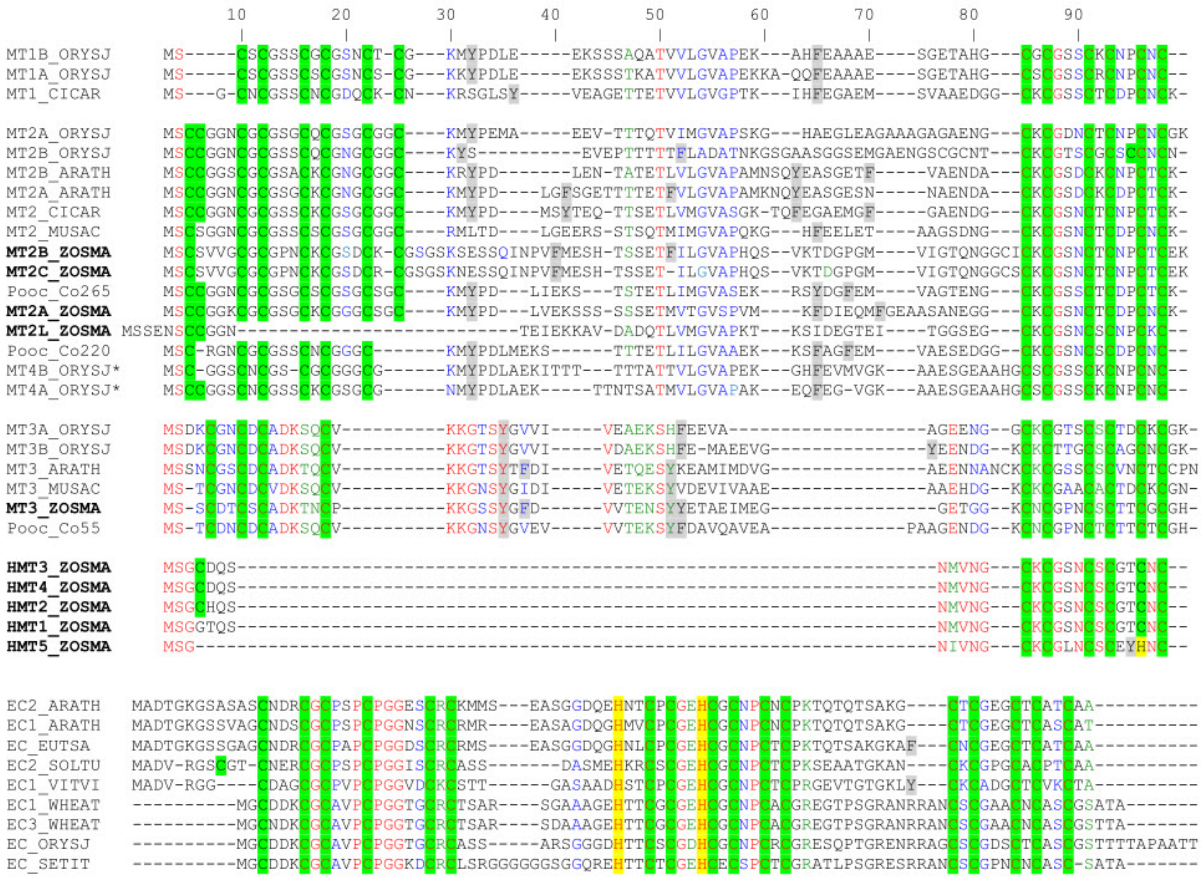
Supplementary Figure 6.1. Number of genes expressed in five tissues of *Z. marina*. (a) Venn-diagram of genes with expression values (FPKM) higher than 1 are considered as expressed in the tissue. **(b)** Pairwise differential gene expression analysis between tissues. The male flower shows the highest number of differentially expressed genes.



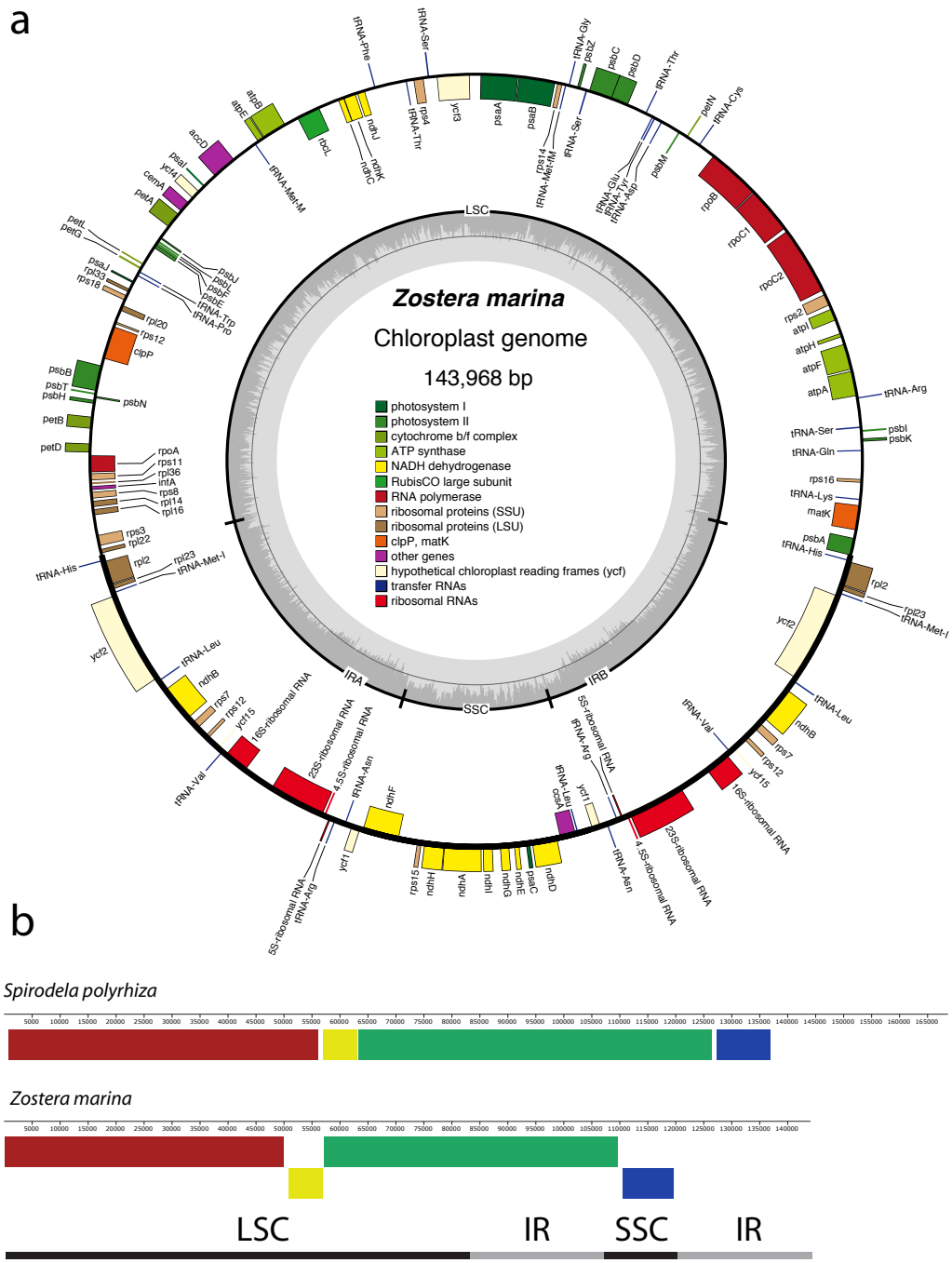
Supplementary Figure 6.2. Circos plot of the 10 largest scaffolds of *Z. marina*. Tracks from outside to inside: GC%, gene density, transposable element (TE) density (density measured in 20 Kbp sliding windows) and gene expression profiles from five tissues (root, leaf, male flower, female flower early and female flower late), presented as log2 FPKM values.



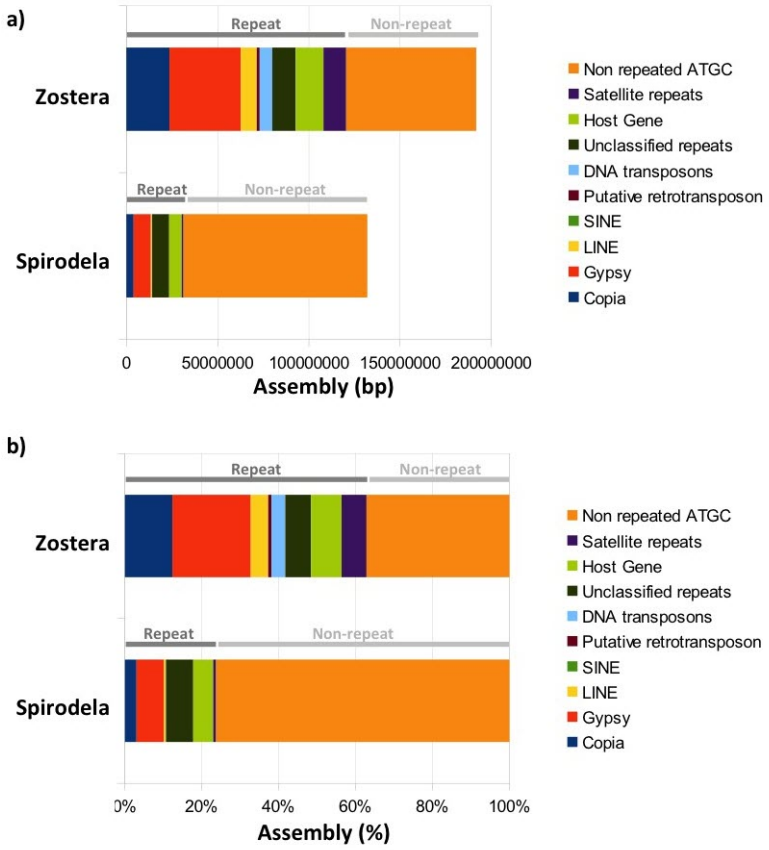
Supplementary Figure 6.3. Potential impact of transposable elements (TEs) on *Z. marina* evolution. (a) Frequency distribution of pairwise sequence identity values between copies of Copia- and Gypsy-type LTR retrotransposons and DNA transposons, and their cognate consensus sequences (younger repeats share higher sequence similarity). Two peaks are detectable for Copia-type elements. **(b)** Distance to the closest TE for the set of *Z. marina* single copy genes and the set of *Z. marina* accessory genes. TE-proximal accessory genes are more frequent than TE-proximal single copy genes. **(c)** Frequency of pairwise sequence identity between accessory gene-proximal Ty3-Gypsy elements and their cognate consensus sequences. A significant number of high identity copies (i.e. putatively young duplicate genes) is observed.



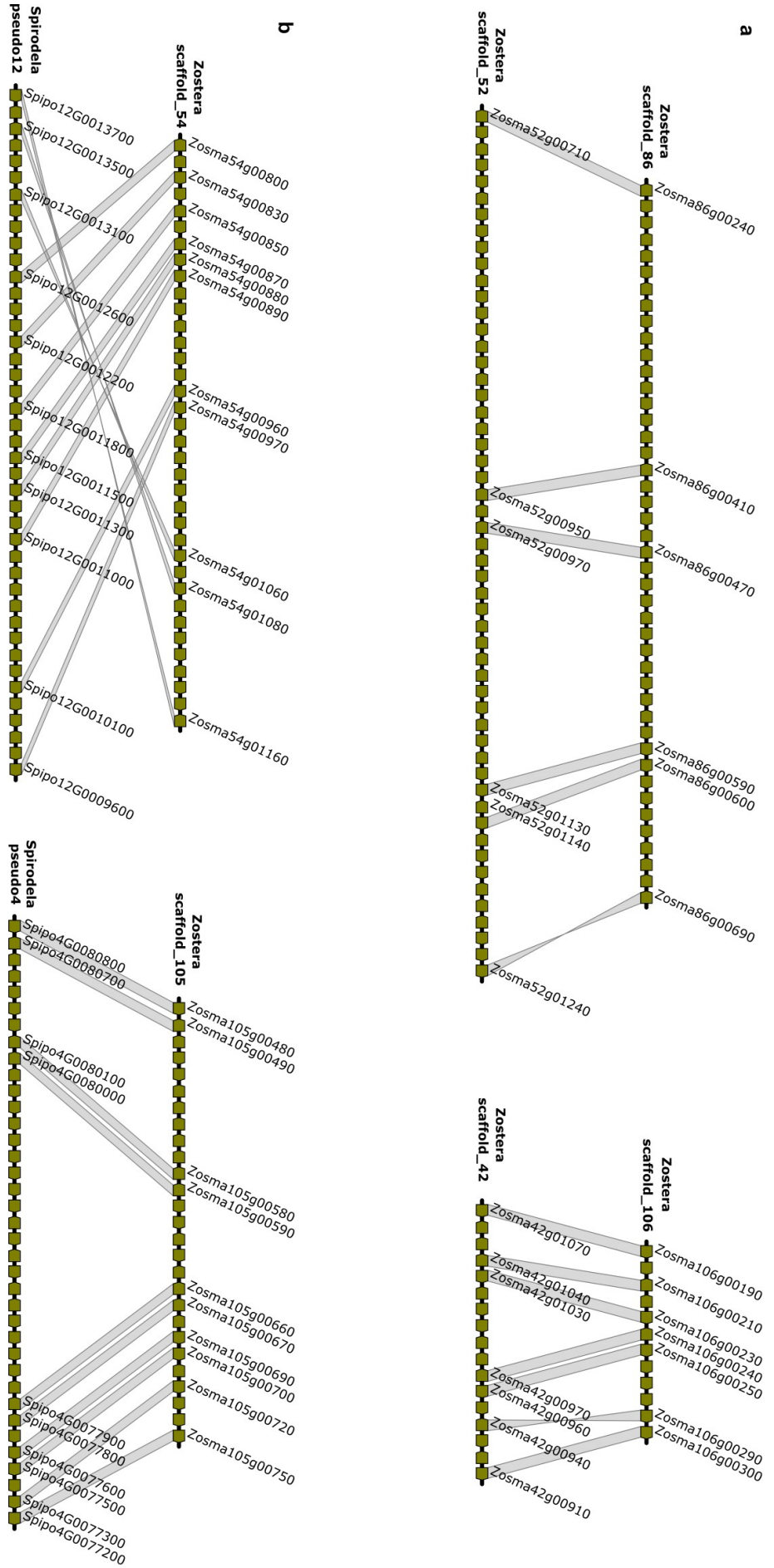
Supplementary Figure 6.4. Alignment of 10 metallothionein (MT) and half-metallothionein (HMT) genes in *Z. marina* as compared with other plants. *Z. marina* genes are highlighted in bold. Alignments were done using ClustalW on the Lyon PBIL web server. The upper alignment is for type 1-3 MTs; the lower alignment is for Type 4 EcMTs where there is no Zostera homolog. Conserved residues are shown in red, and residues in the same amino acid group in blue. The Cys and His residues putatively involved in binding metals are highlighted in green and yellow, respectively. Aromatic amino acids absent in canonical animal MTs are highlighted in grey.



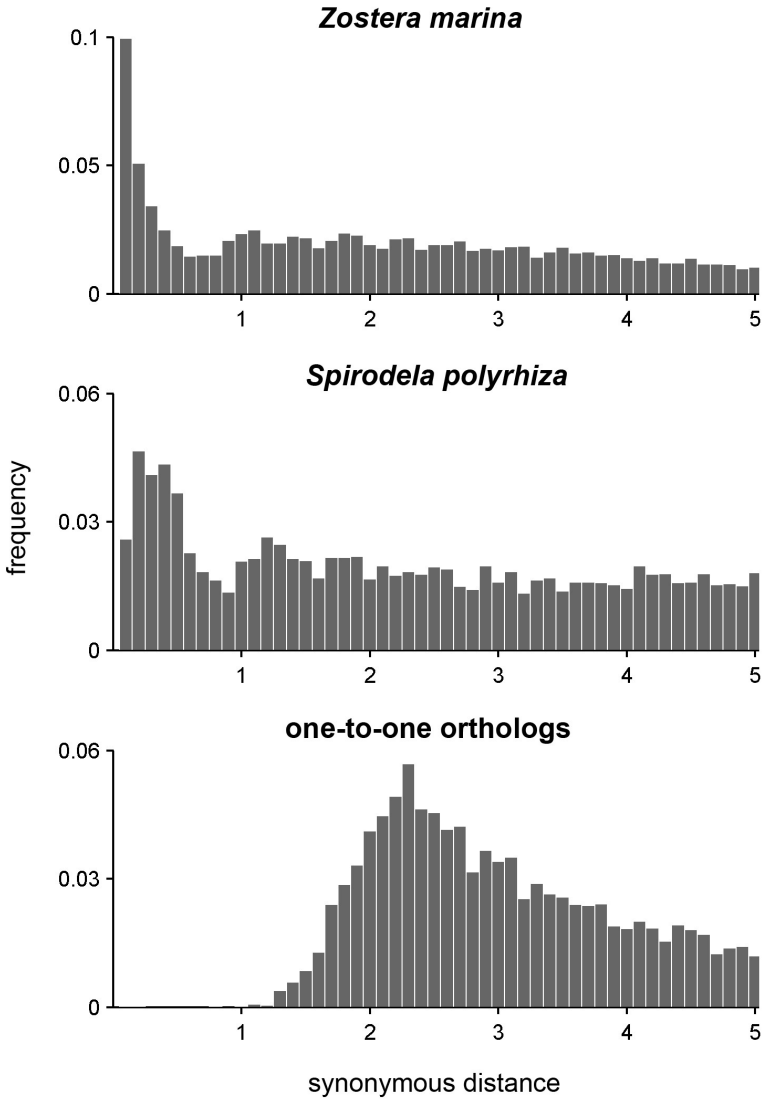
Supplementary Figure 6.5. The *Zostera marina* chloroplast genome and comparison with *Spirodela polyrhiza*. (a) The size of the chloroplast genome differs significantly from that of other species in the Alismatales. A large part of this difference is due to the fact that *Z. marina* has a very short ycf1 gene that codes for the protein Tic214, which is part of the protein translocation system in chloroplasts [504]. The sequence was manually assembled from two contigs into one circular genome. Genes shown on the outside of the circle are transcribed counter-clockwise; genes on the inside are transcribed clockwise. The colour-coded legend in the center shows different classes of genes. Image generated using OrganellarGenomeDRAW [505, `http://ogdraw.mpimp-golm.mpg.de/index.shtml`]. **(b)** The two rows of coloured boxes indicate similar regions in the *S. polyrhiza* (top row) and *Z. marina* (bottom row) chloroplast genomes. Boxes displaced below the row indicate reversed regions. The yellow box indicates the reversed region around the genes rbcL, atpB, atpE and tRNA-Met-M in *Z. marina*, and the blue box represents the main part of SSC that is also reversed compared to *S. polyrhiza*. The thick black and grey line at the bottom of the figure indicates the location of the long single copy region (LSC), short single copy region (SSC) and the two inverted repeats (IRs).



Supplementary Figure 6.6. Repeat-driven genome size difference between *Zostera marina* and *Spirodela polyrhiza*. Repeated and non-repeated fractions of the Zostera and Spirodela genome assemblies as expressed in base pairs **(a)** and in percent of assembly **(b)** are indicated by grey bars above histograms for each species. The different classes of elements composing the repeated fractions are listed with their colour code.



Supplementary Figure 6.7. Examples of syntenic regions within *Z. marina* and between *Z. marina* and *S. polyrhiza*. About 9% of the *Z. marina* genome is covered by such regions of within-genome colinearity. **(a)** Examples of two syntenic regions within *Z. marina*. **(b)** Examples of two syntenic regions between *Z. marina* and *S. polyrhiza*. About 29% of the *Z. marina* genome is collinear with *Spirodela*.



Supplementary Figure 6.8. K_S-based age distributions for *Z. marina*, *Spirodela polyrhiza*, and their one-to-one orthologs. The x-axis shows the synonymous distance (in bins corresponding to a K_S of 0.1), while the y-axis shows the frequency of retained duplicates per bin, for each distribution as indicated on top of the individual panels.

CHAPTER 7

DISCUSSION

I've got an idea – an idea so smart that my head would explode if I even began to know what I'm talking about.

- Peter Griffin

7.1 Introduction

In the previous chapters, we described the annotation of several marine eukaryotic genomes. These organisms, both unicellular and multicellular, small or big genome, contain several peculiar features that stand out in relation to other close-by neighbours and play an important role in the species biology and evolution. Such features include: extremely small genome size, large TE content, genome heterogeneity, outlier chromosomes and repeat introns. These features require specific attention when predicting gene structures. I will go into detail on several of the aforementioned features in order to explain their possible function, origin and evolution, as well as future research possibilities.

7.2 The road ahead

Within this section, I discuss the difficulties and challenges faced in genome projects today as well as possible solutions. Exactly how can we improve the current genome projects?

7.2.1 Plan, plan, plan

Sequencing the human genome took nearly 10 years to complete. Nowadays, the low cost and high throughput of Next-Generation Sequencing techniques has reduced the same work to a few days. With such advances, every scientist can sequence his pet genome. However, deciding on a correct strategy for sequencing and assembly is vital for the downstream analysis. Firstly, what do we want to learn about this species? It is possible to answer many questions on a species' functioning without the need for a genome assembly. Secondly, which libraries are going to be sequenced, which technology, which assembly strategy? A priori information (e.g. genome size; number of repeats) is able to guide such decisions. Sequencing a 100Gb genome on 150bp Illumina libraries for instance, will not result in a decent assembly. You will need several jumping libraries, long sequencing reads, and potentially different genetic or physical maps, in order to produce something that other scientists can use. On top of this, scientists have to be aware that sequencing loads of libraries can confuse the software due to the added data complexity. Deciding on 'what, when and how much' is extremely important.

Thirdly, do we have the capacity to store and analyse all that data? Gathering the data in one place is already rather challenging, requiring enough bandwidth and loads of storage space. Additionally, the computational analysis (assembly, annotation) requires high-memory machines and entire computer clusters. While more scientific groups have access to such infrastructure, it is vital to set aside a proper IT budget within the scope of the project. Finally, communicate a proper end date for specific work packages. It is commonplace in genome projects that initial steps of the genome project (sequencing, assembly, annotation) drag on, which not only decreases the time other collaborators have to analyse the results, but also increases the chance of being scooped on this particular subject. It is always possible to improve, but marginal gains do not always justify the amounts of time invested.

7.2.2 New technologies & new software

Long reads are the inevitable future for genome projects. In assembly, the combination of long reads and optical maps seems the way forward. Currently, the cost, computational requirements and rate of sequencing errors favours the hybrid approach over a pure SMRT assembly. While the technology is maturing fast, the software to cope with this kind of data is not. The limited choice in PacBio assembly software coupled with absurd system requirements (memory usage, specific system environment) means that even now, many genome projects do not take advantage of such technologies. In this thesis, only one project (i.e. *Ostreococcus tauri*) included PacBio data, and it was inevitably discarded because of its abundance of sequencing errors and lack of tools to process the data, requiring the development of custom scripts. The scientific community needs user-friendly software to adopt the new technologies: if nobody is able to use it, even the best software will fade out...

As in assembly, long reads are the future for annotation. The reads are able to span entire genes, providing excellent evidence sources for genome prediction software. Again, we require updated software that is able

to process this type of data. Being able to provide a BAM alignment file would be a great selling point for any prediction software. Nowadays, we have to post-process the alignments and provide the results as filtered pileups (e.g. AUGUSTUS) or extract splice-site junctions (AUGUSTUS, EuGene), just to allow the software to interpret the data.

Finally, the speed at which new technologies are being developed, is so fast that your current genome project strategy might become redundant the next year. Waiting several years before sequencing does result in more reads, less sequencing errors and lower costs. However, are you prepared to hold out so long? This prospect might sound frightening, but it is a very exhilarating thought to work in such a dynamic field. And the next big technology wave is already standing by: the de-centralisation of sequencing power from a few hundred sequencing centres to millions of personalised sequencing machines (e.g. the MinION device).

7.2.3 Updating gene prediction

Methods for gene prediction are well established and are not likely to be updated. Possible improvements include the integration of novel data types (see previous section), optimising training procedures, increasing the user-friendliness of gene prediction pipelines, and reducing the overall time it takes to run a prediction. Predictions can be quite time-consuming, a real problem in a world that sequences loads of genomes every single day *(section 7.2.5)*. One option is to transfer annotations from one already-annotated species onto related species. Another option is the reliance on more and more extrinsic data to reduce the need for *ab initio* training.

The current gene prediction pipelines are already combining different evidence sources to provide accurate predictions. Nevertheless, they all rely on a training set to construct and optimize internal models. The construction of said training set is still open to major improvements. While training sets can be defined manually, which is the case for the genome projects listed in this thesis, it is a time-consuming and laborious task. Automated methods are the way forward. The current state-of-the-art methods will iteratively train themselves starting from an initial 'universal' model (e.g. BRAKER1 [92] and GeneMark-ET [84]). The initial naïve (i.e. untrained) prediction will be used to refine the model parameters. Subsequent iterations of prediction and refinement will shift the model away from its universal nature towards a more species-specific model. This user-friendly automated procedure is also the bottleneck of the method, because the initial universal model introduces a bias in the prediction software from the start. If your species has many properties that deviate from the built-in standard (e.g. weird splice sites, genome heterogeneity), this procedure will have a hard time training the models. In *Micromonas* for instance, the JGI pipeline was unable to accurately predict many gene structures in the CCMP1545 isolate due to the presence of repeat introns (Introner Elements) that were not recognised by the standard models. An alternative, but more time-consuming approach is to run a prediction that relies purely on extrinsic evidence, and select only those genes that have full EST/RNAseq coverage and perfect homologous protein alignments. Stringent criteria and filtering can produce an excellent training set without the need of manual assistance.

Finally, gene prediction has moved beyond identifying protein-coding genes. It catalogues repeat elements, regulatory regions, ncRNA genes and pseudogenes. Gene prediction pipelines today cannot handle such a variety, being restricted mostly to protein-coding genes. Regions containing other types are usually masked to avoid interference from the prediction software, and the respective elements are then added in a post-processing step. The ability to correctly handle and integrate these types would make the entire process more streamlined and less of a patch-up job. This widening scope also implicates that 'older' annotations should be revisited and brought up-to-date.

7.2.4 Standard file formats

Data needs to be structured in such a way that other scientists and software are able to use it. For this purpose, standard file formats were agreed on. In sequencing data, the FASTQ format is well established, while it did encounter its share of problems in relation to alternative header styles e.g. how to set the proper flag to determine the left or right mate within a pair. In assemblies, the outcome is almost always a FASTA file, while many assemblers are also capable of outputting genomic contig data in the ACE format. The results

of annotation however, are less defined. The GENBANK/EMBL format provides a detailed and feature-rich structure, able to handle any annotation description. However, owing to its complexity, many computational scientists shun the format and instead opt for the easier-to-parse tab-delimited General Feature Format (GFF). While the GENBANK/EMBL format is very strict, the GFF format is very flexible. This allows scientists to create their own GFF flavour, which subsequently makes it less universal, requiring different data parsers for each flavour. Such end result is the exact opposite of what standard file formats should achieve. We either need a novel standard (unlikely), or less flexible GFF guidelines.

7.2.5 The annotation struggle: publish or perish?

> I'm gonna go upstairs and alternate between hopeful excitement and suicidal pessimism.
>
> - Chris Griffin

Is it still feasible to put a lot of time (~months) into proper genome annotation when new genomes are being sequenced on a daily basis? Is it worth investing if an 'OK' annotation – or even a low-quality one – will also get published in a high-impact journal? As most scientific budgets rely on the number of publications, this evolution is quite worrisome. The short supply of proper quality assessment tools and the obvious lack of interest by the scientific community to embrace the tools that do exist, don't bode well. Additionally, the time it takes to annotate and the portion of the scientific publication dedicated to it seem to be inversely correlated, creating a highly necessary but under-appreciated job.

For some, an annotation that contains most (~90%) of the genes predicted to an acceptable degree will suffice. Judging by the tsunami of sequenced genomes about to hit us and the time-frame required to properly annotate them, it might be the standard we will have to live with in the future. But do we want that? An annotation is a first step in a genome project and influences all downstream analyses. Compromising on this initial cornerstone might be a risky gamble down the road. van den Berg *et al.* [506] even suggested that each systems biology study should start off with structural and functional gene re-annotation, because it can result in a different knowledge outcome.

Do we have to leave behind our more manual methods and evolve towards semi- or fully-automated pipelines? Of course we do, and such evolution has already started long ago. Yet unlike other annotation groups, we still seem to perform more manual curation, a selling point for many of our collaborators. The manual 'Pierre' approach has led to the discovery of many specific features (special intron classes, metallothioneins, loss of ethylene signalling) and it will be hard to replace with a more automated pipeline. Furthermore, many of these peculiarities were found by accident after manually browsing the genome. A trained human eye – and some luck – can play an important role in genome analysis but translates poorly into automated methods.

We should strive towards a compromise between both views and retain a minimal amount of manual intervention/supervision. Annotation provides the building blocks other scientists use to experiment with. A better annotation produces better downstream results, something all scientists should aspire to.

7.3 Mamiellophyceae & industrial applications

The Mammiellales species described in this thesis combine traits from bacteria (fast doubling time, small genomes), yeasts (ribosome concentration similar to *Saccharomyces cerevisiae* [191]) and plants (photosynthesis). While *Chlamydomonas reinhardtii* is emerging as the leading microalga for industrial application [507], it is unlikely that the prasinophytes will become platforms of mass-production in the biotech industry. However, the biosynthetic capabilities and reduced genome size are excellent selling points for translational research. Reconstructing metabolic pathways [508] and networks [509] provides information that can be translated onto other species [510, 511]. One such example is the *Ostreococcus tauri* nitric oxide synthase enzyme, which confers enhanced plant fitness under adverse growth conditions after transformation [512], which can be of great value when applied to crop species. Another example is the production of asymmetric carotenoids [513].

7.4 The *tauri* reference genome

In *chapter 5* we present a genome update of the first sequenced Mammiellales species, *Ostreococcus tauri*. The goal of this paper was to provide an updated genome sequence and annotation to the scientific community. Through different resequencing approaches, 71.5% of the gaps were filled. However, 477 gaps could not be closed, leaving the *tauri* genome incomplete until another sequencing effort resolves the final hurdles. The annotation however has substantially improved, especially in the outlier chromosomes *(Table 7.1)*. The initial annotation tried to bridge gaps in the genome sequence through the introduction of additional (non-existing) introns. The new annotation has corrected many of such mistakes and brought down the intron count from an absurdly high 6,361 to a reasonable 2,126, which is more in line with other *Ostreococcus* species. The statistics in *Table 7.1* display a decrease in proportion of multi-exon genes, coupled with an increase in validated introns (i.e. introns that have RNAseq junctions as evidence). Its genomic properties *(Table 5.4)* are now much closer to its other *Ostreococcus* relatives e.g. *Ostreococcus lucimarinus*: 7743 protein-coding genes of which 15.6% are multi-exon genes.

Segment	**genes**		**introns**	
	total	multi-exon	total	RNAseq-confirmed
v1 BOC1	345	238 (70.0%)	663	74 (11.2%)
v2 BOC1	232	122 (52.6%)	455	444 (97.6%)
v1 SOC	136	70 (51.5%)	104	5 (4.8%)
v2 SOC	95	10 (10.5%)	11	6 (54.5%)

Table 7.1. Outlier chromosome annotation statistics for the *Ostreococcus tauri* update. v1 = 2006 genome; v2 = updated version.

7.5 The Outlier Chromosomes

The outlier chromosomes Big Outlier Chromosome and Small Outlier Chromosome are present in all sequenced Mamiellales genomes so far (*O. tauri*, *O. lucimarinus*, *O.* sp. RCC809, *M. pusilla* CCMP1545, *M.* sp. RCC299, *B. prasinos*) and were immediately visible due to their low GC content compared to other chromosomes. However, the GC patterns differ substantially between both. In the SOC the GC content is very erratic with many ups and downs along the sequence track, while in BOC the GC content remains quite stable, leaving aside a few high-GC islands *(Supplementary Figure 3.2)*. The low GC patterns and altered gene structures are not restricted to Mamiellales species. *Chlorella variabilis* NC64A for instance, exhibits variations in GC content across its genome that correlate with global expression level, average intron size, and codon usage bias [310]. However, these GC islands are never concentrated into specific chromosomes.

The outlier chromosomes do tend to vary more in size than the other chromosomes [169] and this genome plasticity is reminiscent of that in fungi where it is involved in host-parasite interactions. Many pathogens have a '2-speed genome' with highly dynamic repeat-rich genomic compartments – called 'accessory chromosomes' – that promote accelerated gene evolution [514, 515]. Similarly, *Ostreococcus* outlier chromosomes contain faster evolving genes [318], which suggests that both the Big Outlier Chromosome and Small Outlier Chromosome could be accessory chromosomes that are involved in the rapid evolution of Mamiellales species, in order to adapt to rapid changing environmental conditions.

7.5.1 The Small Outlier Chromosome

In all Mamiellales genomes sequenced so far, the Small Outlier Chromosome is one of the smallest chromosomes present in the genome. While it is the smallest, this chromosome represents one of the biggest puzzles in the Mamiellales genomes.

Firstly, the assembly of the SOCs looks worrisome. The truncated genes and weird gene structures hint towards assembly issues, but when we map the original SANGER reads to the assemblies (data not shown), no specific issues are highlighted. For *Ostreococcus tauri* and NGS data however, we do notice a substantial difference in read coverage: specific regions of the SOC have a higher coverage (up to 2X higher), compared to other chromosomes and to the other SOC regions *(Figures 5.1 and 7.1)*. This could imply that the SOC contains a duplicated region that is collapsed during the assembly process, probably because of it's high sequence identity. This trend does not seem to be continued in other Mamiellales (data not shown), where the SOC (and BOC) coverage does displays more fluctuations than the other chromosomes – probably due to the GC content influencing the sequencing process [516] – but nowhere near the levels we observe for *O. tauri*.

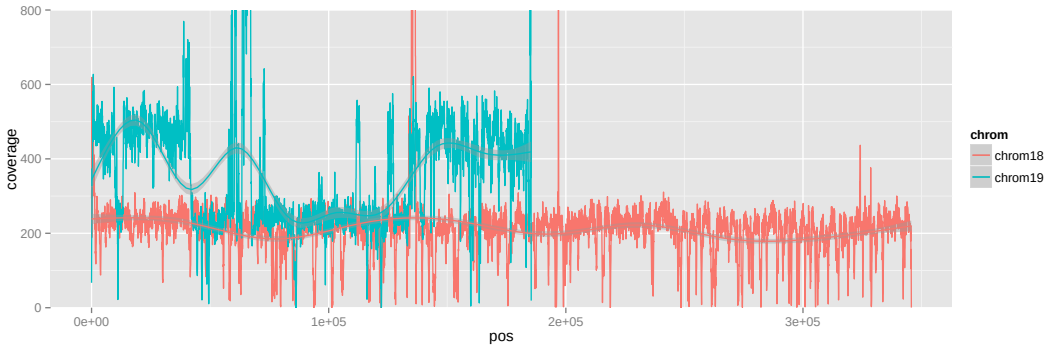


Figure 7.1. gDNA coverage of the Small Outlier Chromosome. For reference, a non-outlier chromosome (chromosome 18) has been added. Average coverage is calculated using a 50nt window. *Ostreococcus tauri* gDNA data source: SRP003551. Mapped with BWA [419]

Secondly, the exact function of the SOC is unknown and rather speculative, mainly due to its strange gene repertoire. Many of the gene models, often truncated, have no known homologs (species-specific genes) [181, 182] *(Figure 3.3 and Table 3.2)*. Those that do are often assigned to specific enriched functional categories such as sugar metabolism (formation of glucoconjugates) [1, 181, 182] and methylation *(Table 3.2)* [1] or encoding for transmembrane proteins that are involved in environment interactions [182, 318]. Because many of these micro-algae are infected by (large) viruses [517, 518], such functional categories can be linked to warding off infections: the SOC serves as an evolutionary hotbed where genes can evolve quickly and allow to change to outer cell wall (sugar metabolism, transmembrane proteins) and detect/eliminate strange virus DNA (methylation). Palenik *et al.* [181] hypothesized that many of the SOC genes have a bacterial origin (horizontal gene transfer, which was later confirmed in *Bathycoccus* [1]. The HGT supplies genes which help the organism disguise itself from phages or grazers through alteration of the cell-surface glycosylation. We can find a similar set-up in the marine cyanobacterium *Prochlorococcus*, with hypervariable genomic islands of non-conserved non-core genes distributed across the genome [519]. Many of these genes are acquired by horizontal gene transfer and code for cell-surface proteins and facilitate coexistence of *Prochlorococcus* and viruses.

Studies have indicated that the outlier chromosomes do indeed play an important role in the host-virus interactions, especially the SOC. In *O.tauri*, strains that have similar SOC sizes tend to be infected by the same viruses, while differences in outlier chromosomes, even in *O. tauri* strains that have identical rDNA sequences, result in altered virus susceptibility patterns [187]. Everything considered, SOC is most likely an accessory chromosome that plays a role in pathogen defence and other environmental interactions.

7.5.2 The Big Outlier Chromosome

To label the Big Outlier Chromosome as an accessory chromosome is more difficult than it is for SOC. To begin with, it does not encompass an entire chromosome, only a part of it (BOC1) [1] *(Supplementary Figure 3.2 and Supplementary Table 3.2)*. While the latter might perfectly well be a fast-evolving accessory chromosome region, it is puzzling why the BOC1 region isn't a stand-alone chromosome like its shorter sibling, SOC.

The BOC1 region contains many vital, highly expressed genes that are of ultimate importance to the Mamiellales species, with a small fraction shared amongst all the sequenced species *(Supplementary Table 3.3)* [1]. If the high gene evolution rate results in any amino acid changes, it is likely to have a severe impact on the organism's functionality, even possibly resulting in cell death. All evidence considered, BOC is not an accessory chromosome, and we suggest a more conservative view in a role as 'sex chromosome' or 'species barrier'.

Sex chromosomes vary widely in sequence composition, gene structure and gene expression compared to the other chromosomes (autosomes). In Mamiellales, BOC1 more than qualifies as it exhibits: 1) extensive gene shuffling *(Figure 3.3, collinearity tracks)*, 2) altered gene structures with a BOC1-specific intron class *(Figures 3.3 and 4.4)*, 3) lower GC content, 4) low recombination rates [520, 521], 5) higher expression values *(Figures 3.3 and 5.1 and Supplementary Figure 3.3)*, and 6) high transposon density [182]. While sexual reproduction is likely to occur in phytoplanktic microalgae [197], there is no known mating-type locus present in any Mamiellales species, and minus/plus haplotypes have not (yet) been established. There aren't many examples of sexual reproduction in unicellular microalgae, mainly because it is a difficult topic to study. Two examples are *Chlamydomonas reinhardtii*, of which the sexual reproductive cycle is well studied *(Figure 7.2)*, and *Nephroselmis olivacea*, where the minus and plus gamete were morphologically similar – 'isogamous' – but behaved differently during the mating process [522]. Over time, many different sex determining mechanisms have evolved [523–525], which makes it likely that the prasinophytes developed their own unique brand of sexual reproductive system.

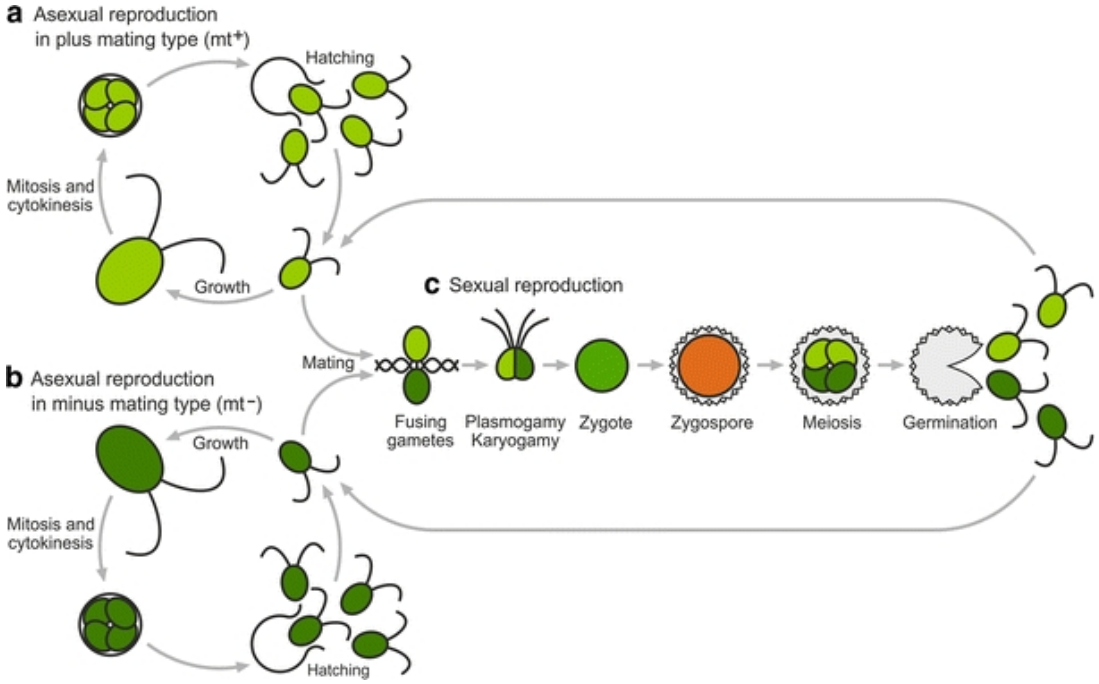


Figure 7.2. Asexual and sexual reproduction in *C. reinhardtii*. (a,b) In favourable conditions, *Chlamydomonas* reproduces asexually. Both mating types have the same asexual life cycle. **(c)** Sexual reproduction is induced by unfavourable conditions. The asexual cells develop into gametes and gametes of opposing mating type fuse together to form a diploid zygote. Meiosis will give rise to four haploid progeny cells, two of each mating type. Source: [526]

The Big Outlier Chromosome might also represent a species barrier, a mechanism of reproductive isolation that helps to maintain species identity. According to the biological species concept, if two populations can't properly mate due to pre-zygotic (e.g. impossible to form the zygote) or post-zygotic (e.g. sterile progeny) mechanisms, they are considered two separate species. In Mamiellales species, the low recombination rates in the BOC1 region will ensure that each species maintains a species-specific identity. Subsequently, when two gametes fuse together to form a new diploid zygote prasinophyte, this zygote can only reproduce if it is able to properly pair both BOC chromosomes or risk creating non-viable aneuploid offspring [181]. Thus, the BOC properties (see previous paragraph) ensure that only gametes originating from the same parental line can produce viable offspring, eliminating inter-strain breeding.

7.5.3 BOC and SOC origin

Although it is tempting to see BOC and SOC as siblings, their many differences – basically everything except for the low GC content – make a common origin virtually impossible. The SOC most likely arose as an accessory chromosome, just like bacteria collect accessory genes on plasmids to gain additional functionality (resistance, compound degradation,...). On these small chromosomes, extensive gene loss and gain mechanisms have run rampant allowing for huge changes in the gene repertoire. This is illustrated by the SOC size differences in *O. tauri* strains [169], as well as the total lack of shared genes between SOCs of different Mamiellales species *(Figure 3.3)*.

In the case of the Big Outlier Chromosome, the scenario is even more puzzling. Where did it come from and why does it accumulate highly-expressed vital genes? Is it possible that this outlier chromosome evolved from a regular chromosome? If we observe the GC content pattern in the BOC1 region, many islands with high GC content are immediately noticeable. This observation is especially true for *Micromonas pusilla* CCMP1545 and *Ostreococcus tauri*. These islands could represent segments of the 'old' BOC that haven't evolved the BOC1 characteristics yet. Or they could represent ancient BOC1 segments that are slowly developing canonical genomic characteristics, thereby loosing BOC1 features.

An alternative scenario involves the integration of the BOC1 ancestor chromosome into another small chromosome, thereby creating this hybrid BOC in the Mamiellales ancestor. In *Micromonas* we can find circumstantial evidence for this integration. If we compare the average density of Introner Elements across the genome, we find a negative correlation between the chromosome length and the number of IEs: the shorter the chromosome, the more IEs occupy it [527]. If we remove the BOC1 region from the Big Outlier Chromosome, the remaining length and IE density is very similar to the smallest chromosomes (with exception of SOC). Another question related to this integration is the functionality of the 'normal' (BOC0) segments on either side. Do they have any kind of functionality in respect to the internal BOC1 region?

7.5.4 Future research

To resolve the SOC issues, it would be helpful to at least have the correct chromosome assembly. We could opt for SMRT sequencing (PacBio), or try to isolate the SOC in order to reduce the sample complexity and sequence this in a more conventional way [528], or a combination of both.

To examine the variation in SOC size, it would be interesting to extract several SOCs from an *O. tauri* population and document the differences, especially in terms of gene content. Which genes are gained or lost? Or is the variation mainly due to variations in the intergenic regions. We could also infect the population with prasinoviruses before starting the SOC analysis to find out if virus genomic segments are being incorporated into the SOC.

A final and important research topic involves the 'spread' of BOC and SOC within the Mamiellales order. Sequencing the genomes of *Mamiella gilva* and *Mantoniella squamata* could provide an answer whether the outlier chromosomes are restricted to the genera *Ostreococcus*, *Micromonas* and *Bathycoccus*, or represent a general Mamiellales feature. Small Outlier Chromosomes are not expected to be found based on viral sensitivity/resistance patterns *(section 3.2.5)*, but the genome sequence itself could resolve this issue. However, even if the two other genera do not possess outlier chromosomes, the most parsimonious solution is the presence of BOC and SOC in the ancestor of Mamiellophyceae, and the subsequent loss in the genera *Mantoniella* and *Mamiella*.

7.6 Introner Elements

7.6.1 The propagation mechanism

Several properties of Introner Elements directly hint at a mechanism that involves transcription and splicing, or in this case: reverse splicing. First and foremost, IEs are fully functioning introns. They have the necessary splicing signals and extrinsic evidence (EST and RNAseq) clearly illustrates their excision from pre-mRNA. Additionally,

they are always found in sense orientation within genes. Both facts indicate an intimate relationship with the transcription and splicing machinery. While the intron transposition mechanism does indeed face several hurdles *(section 2.3.6)*, it is perfectly possible that the mechanism itself is only activated under specific conditions causing 'bursts' of IE propagation, conditions that allow to overcome the aforementioned bottlenecks.

The adversary of intron transposition is mRNA-mediated intron loss. This mechanism is thought to account for the strong 5' position bias in RSIs due to the greater representation of cDNA products arising from the transcript 3' end by the Reverse Transcriptase [529–531]. Such pattern is absent in IEs *(Supplementary Figure 4.7)*. To resolve this issue, Simmons *et al.* [532] proposes a mechanism that is reminiscent of 'spliceosomal retrohoming' [286], but with a twist *(Figure 7.3)*. Instead of targeting double-strand DNA (dsDNA), an 'armed spliceosome' would target single-strand DNA (ssDNA) at R-loops. The nascent RNA strand from an RNA polymerase might pair back with the DNA template strand, displacing an unstable ssDNA R-loop [533], wherein the IE reverse splices. The ability of the spliceosome to target ssDNA directly would be possible [534], but the mechanism – like the others – is still highly hypothetical.

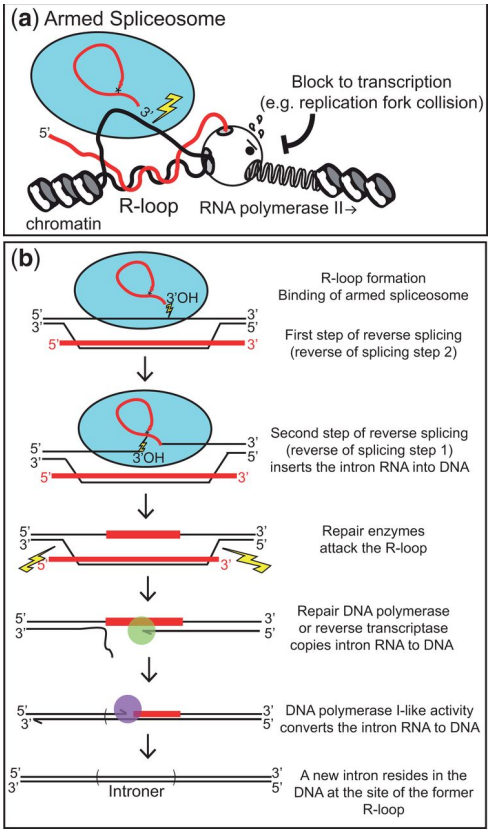


Figure 7.3. Proposed model for IE reverse splicing into ssDNA generated at R-loops. (a) An R-loop forms behind a stalled RNA polymerase II complex by pairing of the RNA transcript (red) and the DNA template strand. **(b)** An armed spliceosome recognises the displaced nontemplate DNA strand and initiates reverse splicing. Subsequent repair mechanisms ensure the integration of the IE in the genome. Source: [532]

While a Big Outlier Chromosome is present in all Mamiellales species, introners are not. Using the parsimony principle, the BOC arose before IEs invaded the *Micromonas* species. But why are there no IEs present in the low GC regions of the outlier chromosomes? What prevents the Introner Elements from invading these regions? The proposed propagation mechanism itself might provide some clues. Low GC content is often associated with lower recombination rates, and recombination of the reverse-transcribed transcript (carrying the reverse spliced IE) is an important step of the intron gain scenario described for Introner Elements (intron transposition). Hence, while transcripts of outlier chromosomes might be invaded by IEs, they never successfully integrate

into the genome due to lack of recombination. Additionally, the proposed mechanisms that work on the DNA level fail to explain the bias towards genic sense regions. Does the spliceosome or IE itself recognise specific proto-splice sites in genic DNA sequences?

The main mechanisms proposed here – whether on the DNA or RNA level [2, 286, 532] – are an almost exact copy of the proposed group II intron mobility mechanisms. For both mechanisms to work, additional proteins would likely be involved. While group II introns operate with the help of intron-encoded proteins, no such arrangements are found in the *Micromonas* or fungal genomes. Only introns belonging to the IE-D class qualify and the function of their IEPs has yet to be established. Furthermore, genes encoding for Reverse Transcriptases similar to those employed by group II introns, could not be found in all IE/ILE-containing genomes [535]. Alternatively, all those genomes do contain genes that code for putatively active RTs of non-LTR retrotransposons which IEs could use [535]. There is a slim chance of course that other uncharacterised genes encode for mobility factors that aid in the propagation of IEs, either by interacting with the IE itself, or through interactions with the spliceosome (allowing efficient splicing and reverse splicing). Further experiments will have to be conducted to investigate the matter. If we take into account the entire history of intron gain research and the few examples it has led to, it will most likely take many more years before we unravel the way IEs propagate.

IE3 sequences pose an additional question. Unlike other IE classes, they maintain a very high intra-class identity. This could imply they are either less prone to degeneration, or they have been replicated more recently. If the latter is true, do the different IE classes replicate independently from each other?

7.6.2 Creating novel spliceosomal introns

Introner Elements degenerate over time. The repertoire of IEs displays a substantial amount of introners that lost their characteristic motifs, up to the point they became a sheer remnant of the IE they once were. Most often, this remnant consists out of a single motif (motif-C), a motif not directly associated with splicing signals (donor, branch point(s), acceptor). The decline in IE identity takes away our ability to properly distinguish 'true' IEs from regular spliceosomal introns. The latter could actually be highly degenerated IEs. Does this mean that IEs are the predecessors of most RSIs? Studies in fungi have shown that ILE degeneration can contribute up to 90% of recent intron gains [286] and ILE gains occur on average 10-fold more frequently than losses [288]. Likewise, IEs could be the ancestors of many of today's *Micromonas* RSIs.

7.6.3 Introners as rybozymes

The current variety of genomes has led to a large compendium of different TE classes and an urgent need for a universal TE classification system. In a recent TE survey by Piégu *et al.* [535], Introner Elements are classified in the 'Group II intron' TE order together with other mobile lariat introns. The propagation mechanisms detailed in previous sections are certainly reminiscent of group II introns. An important property of group II introns is the RNA secondary structure [536]. In *Micromonas*, we could not discern any conserved secondary structures among different IE classes [2], while a recent study claims such secondary structure is indeed present [532]. In fungi, a conserved secondary structure has also been proposed [286]. It is highly likely some form of IE secondary structure exists but the sequence diversity makes it hard to pinpoint conserved segments. It is possible that the lariat structure, the only commonality in all propagation mechanisms, defines their functionality.

7.6.4 Introner Elements as lineage markers

To sample the diversity of phytoplankton communities, molecular phylogenies are established based on the comparison of 18S rDNA sequences or housekeeping genes. Similarly, IEs can be used as markers to distinguish between different phylogenetic clades. In Verhelst *et al.* [2], we distinguish between two different lineages based on the presence or absence of IE-A and IE-C sequences. In Simmons *et al.* [532], additional IE classes are discovered allowing a more broad coverage. IE-C sequences are found in clades A, B and C, and are re-branded 'ABC-IE', while ie-A sequences are only found in clade D and the new classes in clade E (or D-IE and

E-IE accordingly). Similarly, different ILE families are present in different fungal lineages [297]. Thus, the IEs and ILEs are able to distinguish between different clades, but not to the same level as 18S rDNA does.

7.6.5 Future research

Many research opportunities are open with regards to Introner Elements. Firstly, are IEs still mobile? We could grow *Micromonas* in the lab for several hundred and even thousand generations while re-sequencing at specific time intervals. Do we observe Presence/Absence Polymorphisms between the resequenced genomes and the reference? During the time course we could introduce stress factors within the system to provoke introner propagation bursts.

Secondly, can we find additional classes? Simmons *et al.* [532] showed that broader sampling results in the discovery of more classes in more clades. We could focus on the TARA oceans database, extract sequences from Mamiellales species, and try to find novel repeat introns.

Finally, can we identify the components that are related to the propagation mechanism? We could perform a random mutagenesis experiment on *Micromonas* and analyse the cultures through resequencing. If we find cultures where IEs suddenly run rampant, localising the mutation can provide important genetic clues. Instead of a random approach, we can try targeted mutation of genes involved in splicing or reverse transcription, or bring in foreign DNA encoding for specific proteins that could aid IE mobility. Additionally, if we understand how they propagate, can we use Introner Elements as a vector system in other non-*Micromonas* organisms to insert foreign DNA into the genome? As it can be spliced out, it would not interrupt any host genes when integrating.

7.7 Conclusion

Marine eukaryotes are fascinating organisms with many genomic peculiarities that trouble traditional gene prediction software. Whether it is genome heterogeneity, transposable elements, outlier chromosomes, or repeat introns, gene prediction needs proper training in order to accurately document the gene repertoire. Additionally, as detailed in the previous sections, further research is required to delve deeper into their origin, evolution and functionality. I can only hope that the analyses presented in this thesis lead to real-life applications that contribute to the global scientific effort in (i) combating climate change, (ii) waste management, and (iii) developing durable crops.

BIBLIOGRAPHY

1. Moreau, H. *et al.* Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13.** doi:10.1186/gb-2012-13-8-r74, R74 (2012).

2. Verhelst, B., Van de Peer, Y. & Rouzé, P. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biology and Evolution* **5.** doi:10.1093/gbe/evt189, 2393–2401 (2013).

3. Blanc-Mathieu, R. *et al.* An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* **15.** doi:10.1186/1471-2164-15-1103, 1103 (2014).

4. Olsen, J. L. *et al.* Genome re-engineering from land to sea by the seagrass Zostera marina. *Nature.* Accepted (2015).

5. Vandepoele, K. *et al.* pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **15.** doi:10.1111/1462-2920.12174, 2147–2153 (2013).

6. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13.** doi:10.1038/nrg3174, 329–342 (2012).

7. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* **10.** doi:10.1186/1944-3277-10-18, 18 (2015).

8. Chen, C., Khaleel, S. S., Huang, H. & Wu, C. H. Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol Med* **9.** doi:10.1186/1751-0473-9-8, 8 (2014).

9. Bushnell, B. & Rood, J. *BBMap: A Fast, Accurate, Splice-Aware Aligner* 2014. <http://sourceforge.net/projects/bbmap/>.

10. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btu170, 2114–2120 (2014).

11. The Hannon Lab. *The FASTX-Toolkit: FASTQ/A short-reads pre-processing tools* 2009. <http://hannonlab.cshl.edu/fastx_toolkit/>.

12. Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS ONE* **8.** doi:10.1371/journal.pone.0085024, e85024 (Dec. 2013).

13. Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).

14. Patel, R. K. & Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* **7.** doi:10.1371/journal.pone.0030619, e30619 (Feb. 2012).

15. Marais, G., Yorke, J. A. & Zimin, A. QuorUM: An Error Corrector for Illumina Reads. *PLoS ONE* **10.** doi:10.1371/journal.pone.0130821, e0130821 (June 2015).

16. Greenfield, P., Duesing, K., Papanicolaou, A. & Bauer, D. C. Blue: correcting sequencing errors using consensus and context. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btu368, 2723–2732 (2014).

17. Yang, X., Chockalingam, S. P. & Aluru, S. A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics* **14.** doi:10.1093/bib/bbs015, 56–66 (2013).

18. Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing datahigh-throughput sequencing errors and their correction. *Briefings in Bioinformatics.* doi:10.1093/bib/bbv029 (2015).

19. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27.** doi:10.1093/bioinformatics/btr507, 2957–2963 (2011).

20. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13.** doi:10.1186/1471-2105-13-31, 31 (2012).

21. Liu, B. *et al.* COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics* **28.** doi:10.1093/bioinformatics/bts563, 2870–2874 (2012).

22. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btt593, 614–620 (2014).

23. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat Rev Genet* **14.** doi:10.1038/nrg3367, 157–67 (2013).

24. Myers, E. W. *et al.* A Whole-Genome Assembly of Drosophila. *Science* **287.** doi:10.1126/science.287.5461.2196, 2196–2204 (2000).

25. Batzoglou, S. *et al.* ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research* **12.** doi:10.1101/gr.208902, 177–189 (2002).

26. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24.** doi:10.1093/bioinformatics/btn548, 2818–2824 (2008).

27. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotech* **33.** doi:10.1038/nbt.3238, 623–630 (2015).

28. Ilie, L., Haider, B., Molnar, M. & Solis-Oba, R. SAGE: String-overlap Assembly of GEnomes. *BMC Bioinformatics* **15.** doi:10.1186/1471-2105-15-302, 302 (2014).

29. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Research* **19.** doi:10.1101/gr.089532.108, 1117–1123 (2009).

30. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18.** doi:10.1101/gr.074492.107, 821–829 (2008).

31. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1.** doi:10.1186/2047-217X-1-18, 18 (2012).

32. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108.** doi:10.1073/pnas.1017351108, 1513–1518 (2011).

33. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29.** doi:10.1093/bioinformatics/btt476, 2669–2677 (2013).

34. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14.** doi:10.1186/1471-2105-13-S14-S8, S8 (2012).

35. Paulino, D. *et al.* Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* **16.** doi:10.1186/s12859-015-0663-4, 230 (2015).

36. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding preassembled contigs using SSPACE. *Bioinformatics* **27.** doi:10.1093/bioinformatics/btq683, 578–579 (2011).

37. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST–efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15.** doi:10.1186/1471-2105-15-281, 281 (2014).

38. Farrant, G. K. *et al.* WiseScaffolder: an algorithm for the semi-automatic scaffolding of Next Generation Sequencing data. *BMC Bioinformatics* **16.** doi:10.1186/s12859-015-0705-y, 281 (2015).

39. Hunt, M., Newbold, C., Berriman, M. & Otto, T. D. A comprehensive evaluation of assembly scaffolding tools. *Genome Biol* **15.** doi:10.1186/gb-2014-15-3-r42, R42 (2014).

40. Soto-Jimenez, L. M., Estrada, K. & Sanchez-Flores, A. GARM: Genome Assembly, Reconciliation and Merging Pipeline. *Current Topics in Medicinal Chemistry* **14.** doi:10.2174/1568026613666131204110628, 418–424 (2014).

41. Mirebrahim, H., Close, T. J. & Lonardi, S. De novo meta-assembly of ultra-deep sequencing data. *Bioinformatics* **31.** doi:10.1093/bioinformatics/btv226, i9–i16 (2015).

42. Hernandez Wences, A. & Schatz, M. Metassembler: Merging and optimizing de novo genome assemblies. *bioRxiv.* doi:10.1101/016352 (2015).

43. Fierst, J. L. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front Genet* **6.** doi:10.3389/fgene.2015.00220, 220 (2015).

44. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotech* **31.** doi:10.1038/nbt.2727, 1119–1125 (2013).

45. Levy-Sakin, M. & Ebenstein, Y. Beyond sequencing: optical mapping of {DNA} in the age of nanotechnology and nanoscopy. *Current Opinion in Biotechnology* **24.** doi:10.1016/j.copbio.2013.01.009, 690–698 (2013).

46. Muggli, M. D., Puglisi, S. J., Ronen, R. & Boucher, C. Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics* **31.** doi:10.1093/bioinformatics/btv262, i80–i88 (2015).

47. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Meth* **12.** doi:10.1038/nmeth.3454, 780–786 (2015).

48. Faino, L. *et al.* Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome. *MBio* **6.** doi:10.1128/mBio.00936-15, e00936–15 (2015).

49. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotech* **31.** doi:10.1038/nbt.2478, 135–141 (2013).

50. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Molecular Cell* **58.** doi:10.1016/j.molcel.2015.05.004, 586–597 (2015).

51. Utturkar, S. M. *et al.* Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btu391, 2709–2716 (2014).

52. Liao, Y. C., Lin, S. H. & Lin, H. H. Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep* **5.** doi:10.1038/srep08747, 8747 (2015).

53. Pacific Biosciences Community. *Large Genome Assembly with PacBio Long Reads* <https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Large-Genome-Assembly-with-PacBio-Long-Reads>.

54. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23.** doi:10.1016/j.mib.2014.11.014, 110–120. ISSN: 1369-5274 (2015).

55. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10.** doi:10.1038/nmeth.2474, 563–9 (2013).

56. Salmela, L. & Rivals, E. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btu538, 3506–14 (2014).

57. Lee, H. *et al.* Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv.* doi:10.1101/006395 (2014).

58. Hackl, T., Hedrich, R., Schultz, J. & Frster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30.** doi:10.1093/bioinformatics/btu392, 3004–3011 (2014).

59. Ye, C., Hill, C., Ruan, J. & Ma, Z. DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph. *ArXiv e-prints.* arXiv: 1410.2801 [q-bio.GN] (Oct. 2014).

60. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19.** doi:10.1089/cmb.2012.0021, 455–77 (2012).

61. Deshpande, V., Fung, E., Pham, S. & Bafna, V. in *Algorithms in Bioinformatics* (eds Darling, A. & Stoye, J.) 349–363 (Springer Berlin Heidelberg, 2013). ISBN: 978-3-642-40452-8. doi:10.1007/978-3-642-40453-5_27.

62. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15.** doi:10.1186/1471-2105-15-211, 211 (2014).

63. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotech* **30.** doi:10.1038/nbt.2288, 701–707 (2012).

64. English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7.** doi:10.1371/journal.pone.0047768, e47768 (2012).

65. Kosugi, S., Hirakawa, H. & Tabata, S. GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics.* doi:10.1093/bioinformatics/btv465 (2015).

66. Quick, J., Quinlan, A. & Loman, N. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *GigaScience* **3.** doi:10.1186/2047-217X-3-22, 22 (2014).

67. Karlsson, E., Larkeryd, A., Sjodin, A., Forsman, M. & Stenberg, P. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci Rep* **5.** doi:10.1038/srep11996, 11996 (2015).

68. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat Meth* **12.** doi:10.1038/nmeth.3290, 351–356 (2015).

69. Goodwin, S. *et al.* Oxford Nanopore Sequencing, Hybrid Error Correction, and de novo Assembly of a Eukaryotic Genome. *bioRxiv.* doi:10.1101/013490 (2015).

70. Warren, R. L., Vandervalk, B. P., Jones, S. J. M. & Birol, I. LINKS: Scaffolding genome assemblies with kilobase-long nanopore reads. *bioRxiv.* doi:10.1101/016519, 016519+ (Mar. 20, 2015).

71. Madoui, M. A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16.** doi:10.1186/s12864-015-1519-z, 327 (2015).

72. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nat Biotech* **32.** doi:10.1038/nbt.2833, 261–266 (2014).

73. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-3.0* 1996-2010. <http://www.repeatmasker.org>.

74. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6.** doi:10.1186/s13100-015-0041-9, 11 (2015).

75. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21.** doi:10.1093/bioinformatics/bti1018, i351–i358 (2005).

76. Bao, Z. & Eddy, S. R. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research* **12.** doi:10.1101/gr.88502, 1269–1276 (2002).

77. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16.** doi:10.1186/s12859-015-0654-5, 227 (2015).

78. Koch, P., Platzer, M. & Downie, B. R. RepARK - de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research* **42.** doi:10.1093/nar/gku210, e80 (2014).

79. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0* 2008-2015.

80. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering Transposable Element Diversification in <italic>De Novo</italic> Annotation Approaches. *PLoS ONE* **6.** doi:10.1371/journal.pone.0016526, e16526 (Jan. 2011).

81. Hoede, C. *et al.* PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* **9.** doi:10.1371/journal.pone.0091929, e91929 (May 2014).

82. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass - a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp084, 1329–1330 (2009).

83. Hoen, D. *et al.* A call for benchmarking transposable element annotation methods. *Mobile DNA* **6.** doi:10.1186/s13100-015-0044-6, 13 (2015).

84. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* **42.** doi:10.1093/nar/gku557, e119 (Sept. 2014).

85. Testa, A. C., Hane, J. K., Ellwood, S. R. & Oliver, R. P. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* **16.** doi:10.1186/s12864-015-1344-4, 170 (2015).

86. Allen, J. E. & Salzberg, S. L. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21.** doi:10.1093/bioinformatics/bti609, 3596–3603 (2005).

87. Liu, Q., Mackey, A. J., Roos, D. S. & Pereira, F. C. N. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics* **24.** doi:10.1093/bioinformatics/btn004, 597–605 (2008).

88. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9.** doi:10.1186/gb-2008-9-1-r7, R7 (2008).

89. Zickmann, F. & Renard, B. Y. IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy. *BMC Genomics* **16.** doi:10.1186/s12864-015-1315-9, 134 (2015).

90. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12.** doi:10.1186/1471-2105-12-491, 491 (2011).

91. Reid, I. *et al.* SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. *BMC Bioinformatics* **15.** doi:10.1186/1471-2105-15-229, 229 (2014).

92. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics.* doi:10.1093/bioinformatics/btv661 (2015).

93. Hoff, K. & Stanke, M. Current methods for automated annotation of protein-coding genes. *Current Opinion in Insect Science* **7.** doi:10.1016/j.cois.2015.02.008, 8–14 (2015).

94. Bafna, V. & Huson, D. H. The conserved exon method for gene finding. *Proc Int Conf Intell Syst Mol Biol* **8,** 3–12 (2000).

95. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research* **26.** doi:10.1093/nar/26.2.544, 544–548 (1998).

96. Borodovsky, M. & McIninch, J. GENMARK: Parallel gene recognition for both {DNA} strands. *Computers & Chemistry* **17.** doi:10.1016/0097-8485(93)85004-V, 123–133 (1993).

97. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27.** doi:10.1093/nar/27.23.4636, 4636–4641 (1999).

98. Zhang, M. & Marr, T. A weight array method for splicing signal analysis. *Computer applications in the biosciences : CABIOS* **9.** doi:10.1093/bioinformatics/9.5.499, 499–509 (1993).

99. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic {DNA1}. *Journal of Molecular Biology* **268.** doi:10.1006/jmbi.1997.0951, 78–94 (1997).

100. Sleator, R. D. An overview of the current status of eukaryote gene prediction strategies. *Gene* **461.** doi:10.1016/j.gene.2010.04.008, 1–4 (2010).

101. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr Bioinformatics* **3.** doi:10.2174/157489308784340702, 87–97 (2008).

102. Degroeve, S., Saeys, Y., De Baets, B., Rouz, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21.** doi:10.1093/bioinformatics/bti166, 1332–1338 (2005).

103. Eddy, S. R. What is a hidden Markov model? *Nat Biotechnol* **22.** doi:10.1038/nbt1004-1315, 1315–6 (2004).

104. Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4,** 134–42 (1996).

105. Mathé, C., Sagot, M.-F., Schiex, T. & Rouzé, P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* **30.** doi:10.1093/nar/gkf543, 4103–4117 (2002).

106. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19.** doi:10.1093/bioinformatics/btg1080, ii215–ii225 (2003).

107. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24.** doi:10.1093/bioinformatics/btn013, 637–644 (2008).

108. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van de Peer, Y. GenomeView: a next-generation genome browser. *Nucleic Acids Research* **40.** doi:10.1093/nar/gkr995, e12 (2012).

109. Lee, E. *et al.* Web Apollo: a web-based genomic annotation editing platform. *Genome Biology* **14.** doi:10.1186/gb-2013-14-8-r93, R93 (2013).

110. Sterck, L., Billiau, K., Abeel, T., Rouzé, P. & Van de Peer, Y. ORCAE: online resource for community annotation of eukaryotes. *Nat Methods* **9.** doi:10.1038/nmeth.2242, 1041 (Nov. 2012).

111. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3.** doi:10.1186/1471-2105-3-18, 18 (2002).

112. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research* **41.** doi:10.1093/nar/gks1005, D226–D232 (2013).

113. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25.** doi:10.1093/nar/25.5.0955, 955–964 (Mar. 1997).

114. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35.** doi:10.1093/nar/gkm160, 3100–3108 (2007).

115. Chekanova, J. A. Long non-coding {RNAs} and their functions in plants. *Current Opinion in Plant Biology* **27.** doi:10.1016/j.pbi.2015.08.003, 207–216 (2015).

116. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22.** doi:10.1101/gr.132159.111, 1775–1789 (2012).

117. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458.** doi:10.1038/nature07672, 223–227 (2009).

118. Guo, X. *et al.* Advances in long noncoding RNAs: identification, structure prediction and function annotation. *Briefings in Functional Genomics.* doi:10.1093/bfgp/elv022 (2015).

119. Gomes, C. P. *et al.* A Review of Computational Tools in microRNA Discovery. *Front Genet* **4.** doi:10.3389/fgene.2013.00081, 81 (2013).

120. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22.** doi:10.1093/bioinformatics/btl116, 1437–1439 (2006).

121. Zhang, Y. & Sun, Y. *PseudoDomain: Identification of Processed Pseudogenes Based on Protein Domain Classification* in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* doi:10.1145/2382936.2382959 (ACM, Orlando, Florida, 2012), 178–185. ISBN: 978-1-4503-1670-5.

122. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat Meth* **10.** doi:10.1038/nmeth.2340, 221–227 (2013).

123. Tiwari, A. K. & Srivastava, R. A survey of computational intelligence techniques in protein function prediction. *Int J Proteomics* **2014.** doi:10.1155/2014/845479, 845479 (2014).

124. Amar, D. *et al.* Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. *BMC Plant Biol* **14.** doi:10.1186/s12870-014-0329-9, 329 (2014).

125. Käll, L., Krogh, A. & Sonnhammer, E. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **338.** doi:10.1016/j.jmb.2004.03.016, 1027–1036 (2004).

126. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in molecular biology* **396.** doi:10.1007/978-1-59745-515-2_5, 59–70 (2007).

127. Gtz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36.** doi:10.1093/nar/gkn176, 3420–3435 (2008).

128. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research* **42.** doi:10.1093/nar/gkt1253, D231–D239 (2014).

129. Sahraeian, S. M., Luo, K. R. & Brenner, S. E. SIFTER search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Research* **43.** doi:10.1093/nar/gkv461, W141–W147 (2015).

130. Koskinen, P., Törönen, P., Nokso-Koivisto, J. & Holm, L. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* **31.** doi:10.1093/bioinformatics/btu851, 1544–1552 (2015).

131. Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N. & Sundararajan, V. S. Annotation and curation of uncharacterized proteins- challenges. *Front Genet* **6.** doi:10.3389/fgene.2015.00119, 119 (2015).

132. Sivashankari, S. & Shanmughavel, P. Functional annotation of hypothetical proteins - A review. *Bioinformation* **1,** 335–8 (2006).

133. Pellicer, J., Fay, M. F. & Leitch, I. J. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164.** doi:10.1111/j.1095-8339.2010.01072.x, 10–15 (2010).

134. Donmez, N. & Brudno, M. in *Research in Computational Molecular Biology* (eds Bafna, V. & Sahinalp, S.) doi:10.1007/978-3-642-20036-6_5, 38–52 (Springer Berlin Heidelberg, 2011). ISBN: 978-3-642-20035-9.

135. Safonova, Y., Bankevich, A. & Pevzner, P. A. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *J Comput Biol* **22.** doi:10.1089/cmb.2014.0153, 528–45 (2015).

136. Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24.** doi:10.1101/gr.170720.113, 1384–1395 (2014).

137. Huang, S. *et al.* HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res* **22.** doi:10.1101/gr.133652.111, 1581–8 (2012).

138. Bodily, P. M. *et al.* Heterozygous genome assembly via binary classification of homologous sequence. *BMC Bioinformatics* **16 Suppl 7.** doi:10.1186/1471-2105-16-S7-S5, S5 (2015).

139. Pryszcz, L. P. & Gabaldon, T. Redundans: an assembly pipeline for highly heterozygous genomes. *submitted.* <`https://github.com/lpryszcz/redundans`> (2015).

140. Zhou, Q., Su, X., Wang, A., Xu, J. & Ning, K. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE* **8.** doi:10.1371/journal.pone.0060234, e60234 (Apr. 2013).

141. Schmieder, R. & Edwards, R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* **6.** doi:10.1371/journal.pone.0017288, e17288 (Mar. 2011).

142. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* **4.** doi:10.3389/fgene.2013.00237, 237 (2013).

143. Abby, S. S., Touchon, M., De Jode, A., Grimsley, N. & Piganeau, G. Bacteria in Ostreococcus tauri cultures - friends, foes or hitchhikers? *Front Microbiol* **5.** doi:10.3389/fmicb.2014.00505, 505 (2014).

144. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29.** doi:10.1093/bioinformatics/btt086, 1072–1075 (2013).

145. Mikheenko, K., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* (2015).

146. Schatz, M. C. *et al.* Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. *Briefings in Bioinformatics* **14.** doi:10.1093/bib/bbr074, 213–224 (2013).

147. Hunt, M. *et al.* REAPR: a universal tool for genome assembly evaluation. *Genome Biology* **14.** doi:10.1186/gb-2013-14-5-r47, R47 (2013).

148. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23.** doi:10.1093/bioinformatics/btm071, 1061–7 (2007).

149. Simo, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* doi:10.1093/bioinformatics/btv351 (2015).

150. Salzberg, S. L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22.** doi:10.1101/gr.131383.111, 557–567 (Mar. 2012).

151. Levin, L. A. & Le Bris, N. The deep ocean under climate change. *Science* **350.** doi:10.1126/science.aad0126, 766–768 (2015).

152. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304.** doi:10.1126/science.1093857, 66–74 (2004).

153. Rusch, D. B. *et al.* The <italic>Sorcerer II</italic> Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5.** doi:10.1371/journal.pbio.0050077, e77 (Mar. 2007).

154. Bork, P. *et al.* Tara Oceans studies plankton at planetary scale. *Science* **348.** doi:10.1126/science.aac5605, 873 (2015).

155. Bremer, K. Summary of green plant phylogeny and classification. *Cladistics* **1.** doi:10.1111/j.1096-0031.1985.tb00434.x, 369–385 (1985).

156. Sym, S. & Pienaar, R. in *Progress in Phycological Research* 281–376 (Biopress Ltd., 1993).

157. Lewis, L. A. & McCourt, R. M. Green algae and the origin of land plants. *American Journal of Botany* **91.** doi:10.3732/ajb.91.10.1535, 1535–1556 (2004).

158. Marin, B. & Melkonian, M. Mesostigmatophyceae, a New Class of Streptophyte Green Algae Revealed by {SSU} rRNA Sequence Comparisons. *Protist* **150.** doi:10.1016/S1434-4610(99)70041-6, 399–417 (1999).

159. Lemieux, C., Otis, C. & Turmel, M. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. *Nature* **403.** doi:10.1038/35001059, 649–652 (2000).

160. Guillou, L. *et al.* Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155.** doi:10.1078/143446104774199592, 193–214 (2004).

161. Viprey, M., Guillou, L., Ferréol, M. & Vaulot, D. Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environmental Microbiology* **10.** doi:10.1111/j.1462-2920.2008.01602.x, 1804–1822 (2008).

162. Leliaert, F., Verbruggen, H. & Zechman, F. W. Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays* **33.** doi:10.1002/bies.201100035, 683–692 (2011).

163. Leliaert, F. *et al.* Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31.** doi:10.1080/07352689.2011.615705, 1–46 (2012).

164. Lemieux, C., Otis, C. & Turmel, M. Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC Genomics* **15.** doi:10.1186/1471-2164-15-857, 857 (2014).

165. Eikrem, W. & Throndsen, J. The ultrastructure of Bathycoccus gen. nov. and B. prasinos sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia* **29.** doi:10.2216/i0031-8884-29-3-344.1, 344–350 (1990).

166. Marin, B. & Melkonian, M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. *Protist* **161.** doi:10.1016/j.protis.2009.10.002, 304–336 (2010).

167. Slapeta, J., Lopez-Garcia, P. & Moreira, D. Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Molecular biology and evolution* **23.** doi:10.1093/molbev/msj001, 23–9 (2006).

168. Foulon, E. *et al.* Ecological niche partitioning in the picoplanktonic green alga Micromonas pusilla: evidence from environmental surveys using phylogenetic probes. *Environmental microbiology* **10.** doi:10.1111/j.1462-2920.2008.01673.x, 2433–43 (2008).

169. Subirana, L. *et al.* Morphology, Genome Plasticity, and Phylogeny in the Genus Ostreococcus Reveal a Cryptic Species, O. mediterraneus sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164.** doi:10.1016/j.protis.2013.06.002, 643–659 (2013).

170. Courties, C. *et al.* Smallest eukaryotic organism. *Nature* **370.** doi:10.1038/370255a0, 255–255 (1994).

171. Chrétiennot-Dinet, M. J. *et al.* A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34.** doi:10.2216/i0031-8884-34-4-285.1, 285–292 (1995).

172. Courties, C. *et al.* PHYLOGENETIC ANALYSIS AND GENOME SIZE OF OSTREOCOCCUS TAURI (CHLOROPHYTA, PRASINOPHYCEAE). *Journal of Phycology* **34.** doi:10.1046/j.1529-8817.1998.340844.x, 844–849 (1998).

173. Keeling, P. J. Ostreococcus tauri: seeing through the genes to the genome. *Trends In Genetics* **23.** doi:10.1016/j.tig.2007.02.008, 151–154 (2007).

174. Cavalier-Smith, T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of botany* **95.** doi:10.1093/aob/mci010, 147–75 (2005).

175. Patrushev, L. I. & Minkevich, I. G. The problem of the eukaryotic genome size. *Biochemistry. Biokhimiia* **73.** doi:10.1134/S0006297908130117, 1519–52 (2008).

176. Finkel, Z. V. *et al.* Phytoplankton in a changing world: cell size and elemental stoichiometry. *Journal of Plankton Research* **32.** doi:10.1093/plankt/fbp098, 119–137 (2010).

177. Clark, J. R., Lenton, T. M., Williams, H. T. P. & Daines, S. J. Environmental selection and resource allocation determine spatial patterns in picophytoplankton cell size. *Limnology and Oceanography* **58.** doi:10.4319/lo.2013.58.3.1008, 1008–1022 (2013).

178. Rodriguez, F. *et al.* Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environ Microbiol* **7.** doi:10.1111/j.1462-2920.2005.00758.x, 853–9 (2005).

179. Six, C. *et al.* Contrasting photoacclimation costs in ecotypes of the marine eukaryotic picoplankter Ostreococcus. *Limnology and Oceanography* **53.** doi:10.4319/lo.2008.53.1.0255, 255–265 (2008).

180. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *The ISME journal* **5.** doi:10.1038/ismej.2010.209, 1095–107 (2011).

181. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104.** doi:10.1073/pnas.0611046104, 7705–10 (2007).

182. Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences of the United States of America* **103.** doi:10.1073/pnas.0604795103, 11647–11652 (2006).

183. Derelle, E. *et al.* DNA LIBRARIES FOR SEQUENCING THE GENOME OF OSTREOCOCCUS TAURI (CHLOROPHYTA, PRASINOPHYCEAE): THE SMALLEST FREE-LIVING EUKARYOTIC CELL. *Journal of Phycology* **38.** doi:10.1046/j.1529-8817.2002.02021.x, 1150–1156 (2002).

184. Derelle, E. *et al.* Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga Ostreococcus tauri. *PLoS One* **3.** doi:10.1371/journal.pone.0002250, e2250 (2008).

185. Weynberg, K. D., Allen, M. J., Ashelford, K., Scanlan, D. J. & Wilson, W. H. From small hosts come big viruses: the complete genome of a second Ostreococcus tauri virus, OtV-1. *Environmental microbiology* **11.** doi:10.1111/j.1462-2920.2009.01991.x, 2821–39 (2009).

186. Weynberg, K. D., Allen, M. J., Gilg, I. C., Scanlan, D. J. & Wilson, W. H. Genome sequence of Ostreococcus tauri virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. *Journal of virology* **85.** doi:10.1128/JVI.02131-10, 4520–9 (2011).

187. Clerissi, C., Desdevises, Y. & Grimsley, N. Prasinoviruses of the Marine Green Alga Ostreococcus tauri Are Mainly Species Specific. *Journal of virology* **86.** doi:10.1128/JVI.07221-11, 4611–9 (2012).

188. Moreau, H. *et al.* Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. *Journal of virology* **84.** doi:10.1128/JVI.01123-10, 12555–63 (2010).

189. Filee, J. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Front Microbiol* **6.** doi:10.3389/fmicb.2015.00593, 593 (2015).

190. Van Ooijen, G., Knox, K., Kis, K., Bouget, F. Y. & Millar, A. J. Genomic transformation of the picoeukaryote Ostreococcus tauri. *J Vis Exp.* doi:10.3791/4074, e4074 (2012).

191. Henderson, G. P., Gan, L. & Jensen, G. J. 3-D ultrastructure of O. tauri: electron cryotomography of an entire eukaryotic cell. *PLoS One* **2.** doi:10.1371/journal.pone.0000749, e749 (2007).

192. Karpov, P. A., Raevskii, A. V. & Blium Ia, B. Bioinformatic search for plant homologs of protein kinase BUB1–the keypoint of mitotic spindle assembly. *TSitologiia i genetika* **44,** 57–69 (2010).

193. Hindle, M. M. *et al.* The reduced kinome of Ostreococcus tauri: core eukaryotic signalling components in a tractable model species. *BMC Genomics* **15.** doi:10.1186/1471-2164-15-640, 640 (2014).

194. Gan, L., Ladinsky, M. S. & Jensen, G. J. Organization of the smallest eukaryotic spindle. *Curr. Biol.* **21.** doi:10.1016/j.cub.2011.08.021, 1578–1583 (Sept. 2011).

195. Gan, L., Ladinsky, M. & Jensen, G. Chromatin in a marine picoeukaryote is a disordered assemblage of nucleosomes. *Chromosoma* **122.** doi:10.1007/s00412-013-0423-z, 377–386 (2013).

196. Guo, L. & Yang, G. Predicting the reproduction strategies of several microalgae through their genome sequences. *Journal of Ocean University of China.* doi:10.1007/s11802-014-2442-7, 1–12 (2014).

197. Grimsley, N., Péquin, B., Bachy, C., Moreau, H. & Piganeau, G. Cryptic Sex in the Smallest Eukaryotic Marine Green Alga. *Molecular Biology and Evolution* **27.** doi:10.1093/molbev/msp203, 47–54 (2010).

198. Vandepoele, K. *et al.* Genome-Wide Analysis of Core Cell Cycle Genes in Arabidopsis. *The Plant Cell* **14.** doi:10.1105/tpc.010445, 903–916 (2002).

199. Robbens, S. *et al.* Genome-wide analysis of core cell cycle genes in the unicellular green alga Ostreococcus tauri. *Molecular biology and evolution* **22.** doi:10.1093/molbev/msi044, 589–97 (2005).

200. Farinas, B. *et al.* Natural synchronisation for the study of cell division in the green unicellular alga Ostreococcus tauri. *Plant molecular biology* **60.** doi:10.1007/s11103-005-4066-1, 277–92 (2006).

201. Moulager, M. *et al.* Light-dependent regulation of cell division in Ostreococcus: evidence for a major transcriptional input. *Plant Physiol* **144.** doi:10.1104/pp.107.096149, 1360–9 (2007).

202. Corellou, F. *et al.* Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *Plant Cell* **21.** doi:10.1105/tpc.109.068825, 3436–49 (2009).

203. Morant, P. E. *et al.* A robust two-gene oscillator at the core of Ostreococcus tauri circadian clock. *Chaos* **20.** doi:10.1063/1.3530118, 045108 (2010).

204. Thommen, Q. *et al.* Robustness of circadian clocks to daylight fluctuations: hints from the picoeucaryote Ostreococcus tauri. *PLoS computational biology* **6.** doi:10.1371/journal.pcbi.1000990, e1000990 (2010).

205. O'Neill, J. S. *et al.* Circadian rhythms persist without transcription in a eukaryote. *Nature* **469.** doi:10.1038/nature09654, 554–8 (2011).

206. Bouget, F. Y. *et al.* Transcriptional versus non-transcriptional clocks: A case study in Ostreococcus. *Mar Genomics* **14.** doi:10.1016/j.margen.2014.01.004, 17–22 (2014).

207. Six, C., Worden, A. Z., Rodriguez, F., Moreau, H. & Partensky, F. New insights into the nature and phylogeny of prasinophyte antenna proteins: Ostreococcus tauri, a case study. *Molecular biology and evolution* **22.** doi:10.1093/molbev/msi220, 2217–30 (2005).

208. Swingley, W. D. *et al.* Characterization of photosystem I antenna proteins in the prasinophyte Ostreococcus tauri. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1797.** doi:10.1016/j.bbabio.2010.04.017, 1458–1464 (2010).

209. Heijde, M. *et al.* Characterization of two members of the cryptochrome/photolyase family from Ostreococcus tauri provides insights into the origin and evolution of cryptochromes. *Plant, cell & environment* **33.** doi:10.1111/j.1365-3040.2010.02168.x, 1614–26 (2010).

210. Sanchez-Ferandin, S., Leroy, F., Bouget, F. Y. & Joux, F. A New, Sensitive Marine Microalgal Recombinant Biosensor using Luminescence Monitoring for the Toxicity Testing of Antifouling Biocides. *Appl Environ Microbiol* **79.** doi:10.1128/AEM.02688-12, 631–638 (2013).

211. Zhang, S. Y., Sun, G. X., Yin, X. X., Rensing, C. & Zhu, Y. G. Biomethylation and volatilization of arsenic by the marine microalgae Ostreococcus tauri. *Chemosphere* **93.** doi:10.1016/j.chemosphere.2013.04.063, 47–53 (2013).

212. Knight-Jones, E. W. & Walne, P. R. Chromulina pusilla Butcher, a Dominant Member of the Ultraplankton. *Nature* **167.** doi:10.1038/167445a0, 445–446 (1951).

213. Manton, I. Electron microscopical observations on a very small flagellate: the problem of Chromulina pusilla Butcher. *Journal of the Marine Biological Association of the United Kingdom* **38.** doi:10.1017/S0025315400006111, 319–333 (1959).

214. Manton, I. & Parke, M. Further observations on small green flagellates with special reference to possible relatives of Chromulina pusilla Butcher. *Journal of the Marine Biological Association of the United Kingdom* **39.** doi:10.1017/S0025315400013321, 275–298 (02 June 1960).

215. Omoto, C. K. & Witman, G. B. Functionally significant central-pair rotation in a primitive eukaryotic flagellum. *Nature* **290,** 708–710 (Apr. 1981).

216. Mayer, J. A. Viral infection in marine Prasinophycean alga, Micromonas pusilla. *Journal of Phycology,* 229–301 (1977).

217. Cottrell, M. T. & Suttle, C. A. Genetic Diversity of Algal Viruses Which Lyse the Photosynthetic Picoflagellate Micromonas pusilla (Prasinophyceae). *Applied and environmental microbiology* **61,** 3088–91 (1995).

218. Brussaard, C. P., Noordeloos, A. A., Sandaa, R. A., Heldal, M. & Bratbak, G. Discovery of a dsRNA virus infecting the marine photosynthetic protist Micromonas pusilla. *Virology* **319.** doi:10.1016/j.virol.2003.10.033, 280–91 (2004).

219. Not, F. *et al.* A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Applied and environmental microbiology* **70.** doi:10.1128/AEM.70.7.4064-4072.2004, 4064–72 (2004).

220. Throndsen, J. & Kristiansen, S. Micromonas pusilla (Prasinophyceae) as part of pico-and nanoplankton communities of the Barents Sea. *Polar Research* **10.** doi:10.1111/j.1751-8369.1991.tb00646.x, 201–208 (1991).

221. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324.** doi:10.1126/science.1167222, 268–272 (Apr. 2009).

222. Monier, A., Sudek, S., Fast, N. M. & Worden, A. Z. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *The ISME journal* **7.** doi:10.1038/ismej.2013.70, 1764–1774 (2013).

223. Cox, P. A. & Humphries, C. Hydrophilous pollination and breeding system evolution in seagrasses: a phylogenetic approach to the evolutionary ecology of the Cymodoceaceae. *Botanical Journal of the Linnean Society* **113.** doi:10.1006/bojl.1993.1071, 217–226 (1993).

224. Les, D. H., Cleland, M. A. & Waycott, M. Phylogenetic studies in Alismatidae, II: evolution of marine angiosperms (seagrasses) and hydrophily. *Systematic Botany* **22.** doi:10.2307/2419820, 443–463 (1997).

225. Rollon, R. N., de Ruyter van Steveninck, E. D. & van Vierssen, W. Spatio-temporal variation in sexual reproduction of the tropical seagrass Enhalus acoroides (L.f.) Royle in Cape Bolinao, {NW} Philippines. *Aquatic Botany* **76.** doi:10.1016/S0304-3770(03)00070-6, 339–354 (2003).

226. Ducker, S. C. & Knox, R. B. Submarine pollination in seagrasses. *Nature* **263.** doi:10.1038/263705a0, 705–706 (1976).

227. Petersen, C. G. J. *Fiskenes biologiske forhold Holbaek Fjord* 1890.

228. Petersen, C. G. J. Om Baendeltangens (Zostera marina) Aars-Produktion i de danske Farvande. *Mindeskr Japetus Steenstrups Fods* **9,** 1–20 (1913).

229. Green, E. P. & Short, F. T. *World Atlas of Seagrasses* ISBN: 9780520240476 (University of California Press, 2003).

230. Backman, T. & Barilotti, D. Irradiance reduction: Effects on standing crops of the eelgrass Zostera marina in a coastal lagoon. *Marine Biology* **34.** doi:10.1007/BF00390785, 33–40. ISSN: 0025-3162 (1976).

231. Short, F. T. & Neckles, H. A. The effects of global climate change on seagrasses. *Aquatic Botany* **63.** doi:10.1016/S0304-3770(98)00117-X, 169–196 (1999).

232. Nejrup, L. B. & Pedersen, M. F. Effects of salinity and water temperature on the ecological performance of Zostera marina. *Aquatic Botany* **88.** doi:10.1016/j.aquabot.2007.10.006, 239–246 (2008).

233. Boström, C. *et al.* Distribution, structure and function of Nordic eelgrass (Zostera marina) ecosystems: implications for coastal management and conservation. *Aquatic Conservation: Marine and Freshwater Ecosystems* **24.** doi:10.1002/aqc.2424, 410–434 (2014).

234. Haugen, P., Simon, D. M. & Bhattacharya, D. The natural history of group I introns. *Trends Genet* **21.** doi:10.1016/j.tig.2004.12.007, 111–9 (2005).

235. Hausner, G., Hafez, M. & Edgell, D. R. Bacterial group I introns: mobile RNA catalysts. *Mob DNA* **5.** doi:10.1186/1759-8753-5-8, 8 (2014).

236. Zimmerly, S. & Semper, C. Evolution of group II introns. *Mob DNA* **6.** doi:10.1186/s13100-015-0037-5, 7 (2015).

237. Wank, H., SanFilippo, J., Singh, R. N., Matsuura, M. & Lambowitz, A. M. A Reverse Transcriptase/Maturase Promotes Splicing by Binding at Its Own Coding Segment in a Group {II} Intron {RNA}. *Molecular Cell* **4.** doi:10.1016/S1097-2765(00)80371-8, 239–250 (1999).

238. Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* **3.** doi:10.1101/cshperspect.a003616, a003616 (2011).

239. Christopher, D. A. & Hallick, R. B. Euglena gracilis chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. *Nucleic Acids Research* **17.** doi:10.1093/nar/17.19.7591, 7591–7608 (1989).

240. Yoshihisa, T. Handling tRNA introns, archaeal way and eukaryotic way. *Front Genet* **5.** doi:10.3389/fgene.2014.00213, 213 (2014).

241. Lopes, R. R. S., Kessler, A. C., Polycarpo, C. & Alfonzo, J. D. Cutting, dicing, healing and sealing: the molecular surgery of tRNA. *Wiley Interdisciplinary Reviews: RNA* **6.** doi:10.1002/wrna.1279, 337–349 (2015).

242. Shefer, K., Sperling, J. & Sperling, R. The Supraspliceosome - A Multi-Task Machine for Regulated Pre-mRNA Processing in the Cell Nucleus. *Comput Struct Biotechnol J* **11.** doi:10.1016/j.csbj.2014.09.008, 113–22 (2014).

243. Wahl, M. C. & Luhrmann, R. SnapShot: Spliceosome Dynamics I. *Cell* **161.** doi:10.1016/j.cell.2015.05.050, 1474–1474 e1 (2015).

244. Wahl, M. C. & Luhrmann, R. SnapShot: Spliceosome Dynamics II. *Cell* **162.** doi:10.1016/j.cell.2015.06.061, 456–456 e1 (2015).

245. Wahl, M. C. & Luhrmann, R. SnapShot: Spliceosome Dynamics III. *Cell* **162.** doi:10.1016/j.cell.2015.07.033, 690–690 e1 (2015).

246. Sharp, P. A. & Burge, C. B. Classification of introns: U2-type or U12-type. *Cell* **91.** doi:10.1016/S0092-8674(00)80479-1, 875–9 (1997).

247. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* **74,** 3171–3175 (1977).

248. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12.** doi:10.1016/0092-8674(77)90180-5, 1–8 (1977).

249. Perler, F. *et al.* The evolution of genes: the chicken preproinsulin gene. *Cell* **20.** doi:10.1016/0092-8674(80)90641-8, 555–66 (1980).

250. Gilbert, W. & Glynias, M. On the ancient nature of introns. *Gene* **135.** doi:10.1016/0378-1119(93)90058-B, 137–44 (1993).

251. Roy, S. W. Recent evidence for the exon theory of genes. *Genetica* **118.** doi:10.1023/A:1024190617462, 251–66 (2003).

252. Jr, J. M. L. The recent origins of spliceosomal introns revisited. *Current Opinion in Genetics & Development* **8.** doi:10.1016/S0959-437X(98)80031-2, 637–648 (1998).

253. Lynch, M. & Richardson, A. O. The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12.** doi:10.1016/S0959-437X(02)00360-X, 701–10 (2002).

254. Belshaw, R. & Bensasson, D. The rise and falls of introns. *Heredity* **96.** doi:10.1038/sj.hdy.6800791, 208–213 (2006).

255. Nguyen, H. D., Yoshihama, M. & Kenmochi, N. Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evolutionary Biology* **6.** doi:10.1186/1471-2148-6-69, 69 (2006).

256. Jeffares, D. C., Mourier, T. & Penny, D. The biology of intron gain and loss. *Trends in Genetics* **22.** doi:10.1016/j.tig.2005.10.006, 16–22 (2006).

257. Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature reviews. Genetics* **7.** doi:10.1038/nrg1807, 211–21 (2006).

258. Carmel, L., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Research* **17.** doi:10.1101/gr.6438607, 1034–1044 (2007).

259. Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group II intron lariat. *Nature* **514.** doi:10.1038/nature13790, 193–197 (2014).

260. Madhani, H. D. snRNA catalysts in the spliceosome's ancient core. *Cell* **155.** doi:10.1016/j.cell.2013.11.022, 1213–5 (2013).

261. Sharp, P. "Five easy pieces". *Science* **254.** doi:10.1126/science.1948046, 663 (1991).

262. Glanz, S. & Kük, U. Trans-splicing of organelle introns - a detour to continuous RNAs. *BioEssays* **31.** doi:10.1002/bies.200900036, 921–934 (2009).

263. Amini, Z. N., Olson, K. E. & Müller, U. F. Spliceozymes: Ribozymes that Remove Introns from Pre-mRNAs in Trans. *PLoS ONE* **9.** doi:10.1371/journal.pone.0101932, e101932 (July 2014).

264. Jacobs, J. *et al.* Identification of a Chloroplast Ribonucleoprotein Complex Containing Trans-splicing Factors, Intron RNA, and Novel Components. *Molecular & Cellular Proteomics* **12.** doi:10.1074/mcp.M112.026583, 1912–1925 (2013).

265. Martin, W. & Koonin, E. V. Introns and the origin of nucleuscytosol compartmentalization. *Nature* **440.** doi:10.1038/nature04531, 41–45 (2006).

266. Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3** (2015).

267. Chalamcharla, V. R., Curcio, M. J. & Belfort, M. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev* **24.** doi:10.1101/gad.1905010, 827–36 (2010).

268. Doolittle, W. F. The trouble with (group II) introns. *Proceedings of the National Academy of Sciences* **111.** doi:10.1073/pnas.1405174111, 6536–6537 (2014).

269. CAVALIER-SMITH, T. The Origin of Eukaryote and Archaebacterial Cells. *Annals of the New York Academy of Sciences* **503.** doi:10.1111/j.1749-6632.1987.tb40596.x, 17–54 (1987).

270. Carmel, L. & Chorev, M. The function of introns. *Frontiers in Genetics* **3.** doi:10.3389/fgene.2012.00055 (2012).

271. Gallegos, J. E. & Rose, A. B. The enduring mystery of intron-mediated enhancement. *Plant Science* **237.** doi:10.1016/j.plantsci.2015.04.017, 8–15 (2015).

272. Amit, M. *et al.* Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Reports* **1.** doi:10.1016/j.celrep.2012.03.013, 543–556.

273. Wu, J. *et al.* Systematic analysis of intron size and abundance parameters in diverse lineages. *Science China Life Sciences* **56.** doi:10.1007/s11427-013-4540-y, 968–974 (2013).

274. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA* **4.** doi:10.1002/wrna.1140, 49–60 (2013).

275. Hedberg, A. & Johansen, S. D. Nuclear group I introns in self-splicing and beyond. *Mob DNA* **4.** doi:10.1186/1759-8753-4-17, 17 (2013).

276. Stoddard, B. L. Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mob DNA* **5.** doi:10.1186/1759-8753-5-7, 7 (2014).

277. Martinez-Abarca, F., Barrientos-Duran, A., Fernandez-Lopez, M. & Toro, N. The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Research* **32.** doi:10.1093/nar/gkh616, 2880–2888 (2004).

278. Cousineau, B., Lawrence, S., Smith, D. & Belfort, M. Retrotransposition of a bacterial group II intron. *Nature* **404.** doi:10.1038/35010029, 1018–21 (2000).

279. Dickson, L. *et al.* Retrotransposition of a yeast group II intron occurs by reverse splicing directly into ectopic DNA sites. *Proceedings of the National Academy of Sciences* **98.** doi:10.1073/pnas.231494498, 13207–13212 (2001).

280. Toro, N., Jiménez-Zurdo, J. I. & Garcıa-Rodrıguez, F. M. Bacterial group II introns: not just splicing. *FEMS Microbiology Reviews* **31.** doi:10.1111/j.1574-6976.2007.00068.x, 342–358 (2007).

281. Smith, D., Zhong, J., Matsuura, M., Lambowitz, A. M. & Belfort, M. Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes & Development* **19.** doi:10.1101/gad.1345105, 2477–2487 (2005).

282. Roy, S. W. & Irimia, M. Mystery of intron gain: new data and new models. *Trends in Genetics* **25.** doi:10.1016/j.tig.2008.11.004, 67–73 (2009).

283. Yenerall, P. & Zhou, L. Identifying the mechanisms of intron gain: progress and trends. *Biol Direct* **7.** doi:10.1186/1745-6150-7-29, 29 (2012).

284. Tseng, C.-K. & Cheng, S.-C. Both Catalytic Steps of Nuclear Pre-mRNA Splicing Are Reversible. *Science* **320.** doi:10.1126/science.1158993, 1782–1784 (2008).

285. Cohen, N. E., Shen, R. & Carmel, L. The Role of Reverse Transcriptase in Intron Gain and Loss Mechanisms. *Molecular Biology and Evolution* **29.** doi:10.1093/molbev/msr192, 179–186 (2012).

286. vanderBurgt, A., Severing, E., deWit, P. & Collemare, J. Birth of New Spliceosomal Introns in Fungi by Multiplication of Introner-like Elements. *Current Biology* **22.** doi:10.1016/j.cub.2012.05.011, 1260–1265 (2012).

287. Roy, S. W. & Irimia, M. Genome Evolution: Where Do New Introns Come From? *Current Biology* **22.** doi:10.1016/j.cub.2012.05.017, R529–R531 (2012).

288. Collemare, J., van der Burgt, A. & de Wit, P. J. At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi? *Commun Integr Biol* **6.** doi:10.4161/cib.23147, e23147 (2013).

289. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. *Current Biology* **13.** doi:10.1016/S0960-9822(03)00558-X, 1512–1517 (2003).

290. Fedorov, A., Merican, A. F. & Gilbert, W. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences* **99.** doi:10.1073/pnas.242624899, 16128–16133 (2002).

291. Stoltzfus, A., Logsdon, J. M., Palmer, J. D. & Doolittle, W. F. Intron sliding and the diversity of intron positions. *Proceedings of the National Academy of Sciences* **94.** doi:10.1073/pnas.94.20.10739, 10739–10744 (1997).

292. Rogozin, I. B., Lyons-Weiler, J. & Koonin, E. V. Intron sliding in conserved gene families. *Trends in Genetics* **16.** doi:10.1016/S0168-9525(00)02096-5, 430–432 (2000).

293. Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. & Koonin, E. V. Conservation versus parallel gains in intron evolution. *Nucleic Acids Research* **33.** doi:10.1093/nar/gki316, 1741–1748 (2005).

294. Carmel, L., Rogozin, I., Wolf, Y. & Koonin, E. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evolutionary Biology* **7.** doi:10.1186/1471-2148-7-192, 192 (2007).

295. Li, W., Tucker, A. E., Sung, W., Thomas, W. K. & Lynch, M. Extensive, recent intron gains in Daphnia populations. *Science* **326.** doi:10.1126/science.1179302, 1260–2 (2009).

296. Li, W., Kuzoff, R., Wong, C. K., Tucker, A. & Lynch, M. Characterization of Newly Gained Introns in Daphnia Populations. *Genome Biology and Evolution* **6.** doi:10.1093/gbe/evu174, 2218–2234 (2014).

297. Collemare, J., Beenen, H. G., Crous, P. W., de Wit, P. J. & van der Burgt, A. Novel Introner-Like Elements in fungi Are Involved in Parallel Gains of Spliceosomal Introns. *PLoS One* **10.** doi:10.1371/journal.pone.0129302, e0129302 (2015).

298. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281.** doi:10.1126/science.281.5374.237, 237–240 (July 1998).

299. Boyce, D., Lewis, M. & Worm, B. Global phytoplankton decline over the past century. *Nature* **466.** doi:10.1038/nature09268, 591–596 (2010).

300. Li, W. Primary productivity of prochlorophytes cyanobacteria, and eucaryotic ultraphytoplankton: measurements from flow cytometric sorting. *Limnol Oceanogr* **39.** doi:10.4319/lo.1994.39.1.0169, 169–175 (1994).

301. Worden, A., Nolan, J. & Palenik, B. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnol Oceanogr* **49.** doi:10.4319/lo.2004.49.1.0168, 168–179 (2004).

302. Peers, G. & Niyogi, K. Pond scum genomics: The genomes of Chlamydomonas and Ostreococcus. *Plant Cell* **20.** doi:10.1105/tpc.107.056556, 502–507 (2008).

303. Johnson, P. & McSieburth, J. In-situ morphology and occurrence of eukaryotic phototrophs of bacterial size in the picoplankton of estuarine and oceanic waters. *J Phycol* **18.** doi:10.1111/j.1529-8817.1982.tb03190.x, 318–327 (1982).

304. Marie, D., Zhu, F., Balague, V., Ras, J. & Vaulot, D. Eukaryotic picoplankton communities of the Mediterranean Sea in summer assessed by molecular approaches (DGGE, TTGE, QPCR). *FEMS Microbiol Ecol* **55.** doi:10.1111/j.1574-6941.2005.00058.x, 403–415 (2006).

305. Monier, A. *et al.* Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ Microbiol* **14.** doi:10.1111/j.1462-2920.2011.02576.x, 162–176 (2011).

306. Treusch, A. *et al.* Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J* **6.** doi:10.1038/ismej.2011.117, 481–492 (2011).

307. Cheung, M., Au, C., Chu, K., Kwan, H. & Wong, C. Composition and genetic diversity of picoeukaryotes in subtropical coastal waters as revealed by 454 pyrosequencing. *ISME J* **4.** doi:10.1038/ismej.2010.26, 1053–1059 (2010).

308. Marie, D., Shi X, L., Rigaut-Jalabert, F. & Vaulot, D. Use of flow cytometric sorting to better assess the diversity of small photosynthetic eukaryotes in the English Channel. *FEMS Microbiol Ecol* **72.** doi:10.1111/j.1574-6941.2010.00842.x, 165–178 (2010).

309. Massana, R., Balague, V., Guillou, L. & Pedros-Alio, C. Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiol Ecol* **50.** doi:10.1016/j.femsec.2004.07.001, 231–243 (2004).

310. Blanc, G. *et al.* The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *The Plant cell* **22.** doi:10.1105/tpc.110.076406, 2943–55 (2010).

311. Merchant, S. *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318.** doi:10.1126/science.1143609, 245–250 (2007).

312. Prochnik, S. *et al.* Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* **329.** doi:10.1126/science.1188800, 223–226 (2010).

313. Piganeau, G., Grimsley, N. & Moreau, H. Genome diversity in the smallest marine photosynthetic eukaryotes. *Res Microbiol* **162.** doi:10.1016/j.resmic.2011.04.005, 570–577 (2011).

314. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19.** doi:10.1101/gr.092759.109, 1639–45 (2009).

315. Zhang, X. The epigenetic landscape of plants. *Science* **320.** doi:10.1126/science.1153996, 489–492 (2008).

316. Wong, S. & Wolfe, K. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet* **37.** doi:10.1038/ng1584, 777–782 (2005).

317. Lee, S., Ni, M., Li, W., Shertz, C. & Heitman, J. The evolution of sex: a perspective from the fungal kingdom. *Microbiol Mol Biol Rev* **74.** doi:10.1128/MMBR.00005-10, 298–340 (2010).

318. Jancek, S., Gourbiere, S., Moreau, H. & Piganeau, G. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes (Chlorophyta, Prasinophyceae). *Molecular biology and evolution* **25.** doi:10.1093/molbev/msn168, 2293–2300 (Nov. 2008).

319. Avrani, S., Wurtzel, O., Sharon, I., Sorek, R. & Lindell, D. Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature* **474.** doi:10.1038/nature10172, 604–608 (2011).

320. Bellec, L., Grimsley, N., Derelle, E., Moreau, H. & Desdevises, Y. Abundance, spatial distribution and genetic diversity of Ostreococcus tauri viruses in two different environments. *Env Microbiol Reports* **2.** doi:10.1111/j.1758-2229.2010.00138.x, 313–321 (2010).

321. Thomas, R. *et al.* Acquisition and maintenance of resistance to viruses in eukaryotic phytoplankton populations. *Env Microbiol* **13.** doi:10.1111/j.1462-2920.2011.02441.x, 1412–1420 (2011).

322. Ragan, M., Harlow, T. & Beiko, R. Do different surrogate methods detect lateral genetic transfer events of different relative ages?. *Trends Microbiol* **14.** doi:10.1016/j.tim.2005.11.004, 4–8 (2006).

323. Kurland, C., Canback, B. & Berg, O. Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA* **100.** doi:10.1073/pnas.1632870100, 9658–9662 (2003).

324. Keeling, P. & Palmer, J. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9.** doi:10.1038/nrg2386, 605–618 (2008).

325. Bowler, C. *et al.* The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456.** doi:10.1038/nature07410, 239–244 (2008).

326. Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324.** doi:10.1126/science.1172983, 1724–1726 (2009).

327. Raymond, J. & Kim, H. Possible role of horizontal gene transfer in the colonization of sea ice by algae. *PLoS One* **7.** doi:10.1371/journal.pone.0035968, e35968 (2012).

328. Monier, A. *et al.* Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res* **19.** doi:10.1101/gr.091686.109, 1441–1449 (2009).

329. pico-PLAZA. an integrative resource for cross-species genome analysis in algae. <http://bioinformatics.psb.ugent.be/pico-plaza/>.

330. Cantarel, B. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* doi:10.1093/nar/gkn663, D233–238 (2009).

331. Jeanneau, C. *et al.* Structure-function analysis of the human sialyltransferase ST3GalI. *J Biol Chem* **279.** doi:10.1074/jbc.M311764200, 13461–13468 (2004).

332. Harduin-Lepers, A., Mollicone, R., Delannoy, P. & Oriol, R. The animal sialyltransferases and sialyltransferase-related genes: a phylogenetic approach. *Glycobiology* **15.** doi:10.1093/glycob/cwi063, 805–817 (2005).

333. Melkonian, M. & Preisig, H. A light and electron microscopic study of Scherffelia dubia, a new member of the scaly green flagellates (Prasinophyceae). *Nord J Bot* **6.** doi:10.1111/j.1756-1051.1986.tb00876.x, 235–256 (1986).

334. Moestrup, O. & Walne, P. Studies on scale morphogenesis in the Golgi apparatus of Pyramimonas tetrarhynchus (Prasinophyceae). *J Cell Sci* **36,** 437–459 (1979).

335. Moestrup, O. Scale structure in Mantoniella squamata, with some comments on the phylogeny of the Prasinophyceae (Chlorophyta). *Phycologia* **29.** doi:10.2216/i0031-8884-29-4-437.1, 437–442 (1990).

336. Melkonian, M., Becker, B. & Becker, D. Scale formation in algae. *J Electron Microscopy Technique* **17.** doi:10.1002/jemt.1060170205, 165–178 (1991).

337. Becker, D. & Melkonian, M. N-linked glycoproteins associated with flagellar scales in a flagellate green alga: characterization of interactions. *Eur J Cell Biol* **57,** 109–116 (1992).

338. Becker, B. Anterograde transport of algal scales through the Golgi complex is not mediated by vesicles. *Trends Cell Biol* **5.** doi:10.1016/S0962-8924(00)89047-9, 305–307 (1995).

339. Al-Khodor, S., Price, C., Kalia, A. & Kwaik, A. Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol* **18.** doi:10.1016/j.tim.2009.11.004, 132–139 (2010).

340. Brayer, K. & Segal, D. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50.** doi:10.1007/s12013-008-9008-5, 111–131 (2008).

341. Winnepenninckx, B., Backeljau, T. & De Wachter, R. Extraction of high molecular weight DNA from molluscs. *Trends Genet* **9,** 407 (1993).

342. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12.** doi:10.1101/gr.208902, 177–189.52 (2002).

343. Schiex, T., Moisan, A. & Rouze, P. EUGENE: an eukaryotic gene finder that combines several sources of evidence. *Lect Notes Comput Sci* **2066.** doi:10.1007/3-540-45727-5_10, 111–125 (2001).

344. Bathycoccus Genome Annotation Database at Ghent University. <http://bioinformatics.psb.ugent.be/webtools/bogas/>.

345. Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25.** doi:10.1093/nar/25.17.3389, 3389–33402 (1997).

346. Proost, S. *et al.* i-ADHoRe 3 0 - fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* **40.** doi:10.1093/nar/gkr955, e11 (2012).

347. Gremme, G., Brendel, V., Sparks, M. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information Software Technol* **47.** doi:10.1016/j.infsof.2005.09.005, 965–978 (2005).

348. Rogers, M. *et al.* A complex and punctate distribution of three eukaryotic genes derived by lateral gene transfer. *BMC Evol Biol* **7.** doi:10.1186/1471-2148-7-89, 89 (2007).

349. Yi, G., Sze, S. & Thon, M. Identifying clusters of functionally related genes in genomes. *Bioinformatics* **23.** doi:10.1093/bioinformatics/btl673, 1053–1060 (2007).

350. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110.** doi:10.1159/000084979, 462–467 (2005).

351. Becher, V., Deymonnaz, A. & Heiber, P. Efficient computation of all perfect repeats in genomic sequences of up to half a gigabyte, with a case study on the human genome. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp321, 1746–1753 (July 2009).

352. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9.** doi:10.1186/1471-2105-9-18, 18 (2008).

353. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37.** doi:10.1093/nar/gkp759, 7002–7013 (Nov. 2009).

354. Kalyanaraman, A. & Aluru, S. Efficient algorithms and software for detection of full-length LTR retrotransposons. *J Bioinform Comput Biol* **4.** doi:10.1109/CSB.2005.31, 197–216 (Apr. 2006).

355. Llorens, C., Futami, R., Bezemer, D. & Moya, A. The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res.* **36.** doi:10.1093/nar/gkm697, 38–46 (Jan. 2008).

356. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp157, 1335–1337 (May 2009).

357. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32.** doi:10.1093/nar/gkh340, 1792–1797 (2004).

358. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59.** doi:10.1093/sysbio/syq010, 307–321 (2010).

359. Irimia, M. & Roy, S. W. Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genetics* **4.** doi:10.1371/journal.pgen.1000148, e1000148 (2008).

360. Wickham, H. *ggplot2: elegant graphics for data analysis* doi:10.1007/978-0-387-98141-3. ISBN: 978-0-387-98140-6. <http://had.co.nz/ggplot2/book> (Springer New York, 2009).

361. Denoeud, F. *et al.* Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330.** doi:10.1126/science.1194167, 1381–5 (2010).

362. Iwata, H. & Gotoh, O. Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics* **12.** doi:10.1186/1471-2164-12-45, 45 (2011).

363. Szafranski, K. *et al.* Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns. *Genome Biol* **8.** doi:10.1186/gb-2007-8-8-r154, R154 (2007).

364. Jaillon, O. *et al.* Translational control of intron splicing in eukaryotes. *Nature* **451.** doi:10.1038/nature06495, 359–62 (2008).

365. Parenteau, J. *et al.* Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147.** doi:10.1016/j.cell.2011.08.044, 320–31 (2011).

366. Sakurai, A. *et al.* On biased distribution of introns in various eukaryotes. *Gene* **300.** doi:10.1016/S0378-1119(02)01035-1, 89–95 (2002).

367. Nielsen, C. B., Friedman, B., Birren, B., Burge, C. B. & Galagan, J. E. Patterns of intron gain and loss in fungi. *PLoS Biol* **2.** doi:10.1371/journal.pbio.0020422, e422 (2004).

368. Sun, S. *et al.* Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39.** doi:10.1093/nar/gkq1102, D546–51 (2011).

369. Bartschat, S. & Samuelsson, T. U12 type introns were lost at multiple occasions during evolution. *BMC Genomics* **11.** doi:10.1186/1471-2164-11-106, 106 (2010).

370. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302.** doi:10.1126/science.1089370, 1401–4 (2003).

371. Lynch, M. The origins of eukaryotic gene structure. *Mol Biol Evol* **23.** doi:10.1093/molbev/msj050, 450–68 (2006).

372. Gilbert, W. Why genes in pieces? *Nature* **271.** doi:10.1038/271501a0, 501 (1978).

373. Rogozin, I. B., Carmel, L., Csuros, M. & Koonin, E. V. Origin and evolution of spliceosomal introns. *Biol Direct* **7.** doi:10.1186/1745-6150-7-11, 11 (2012).

374. Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* **22.** doi:10.1093/molbev/msi091, 1053–66 (2005).

375. Roy, S. W. & Gilbert, W. Complex early genes. *Proc Natl Acad Sci U S A* **102.** doi:10.1073/pnas.0408355101, 1986–91 (2005).

376. Csuros, M., Rogozin, I. B. & Koonin, E. V. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol* **7.** doi:10.1371/journal.pcbi.1002150, e1002150 (2011).

377. Roy, S. W. & Penny, D. A very high fraction of unique intron positions in the intron-rich diatom Thalassiosira pseudonana indicates widespread intron gain. *Mol Biol Evol* **24.** doi:10.1093/molbev/msm048, 1447–57 (2007).

378. Koonin, E. V. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* **1.** doi:10.1186/1745-6150-1-22, 22 (2006).

379. Da Lage, J. L., Binder, M., Hua-Van, A., Janecek, S. & Casane, D. Gene make-up: rapid and massive intron gains after horizontal transfer of a bacterial alpha-amylase gene to Basidiomycetes. *BMC Evolutionary Biology* **13.** doi:10.1186/1471-2148-13-40, 40 (2013).

380. Torriani, S. F., Stukenbrock, E. H., Brunner, P. C., McDonald, B. A. & Croll, D. Evidence for extensive recent intron transposition in closely related fungi. *Current Biology* **21.** doi:10.1016/j.cub.2011.10.041, 2017–22 (2011).

381. Ohm, R. A. *et al.* Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathog* **8.** doi:10.1371/journal.ppat.1003037, e1003037 (2012).

382. Lynch, M. Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A* **99.** doi:10.1073/pnas.092595699, 6118–23 (2002).

383. Cavalier-Smith, T. Selfish DNA and the origin of introns. *Nature* **315.** doi:10.1038/315283b0, 283–4 (1985).

384. Lynch, M. & Kewalramani, A. Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol Biol Evol* **20.** doi:10.1093/molbev/msg068, 563–71 (2003).

385. Hoskins, A. A. & Moore, M. J. The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem Sci* **37.** doi:10.1016/j.tibs.2012.02.009, 179–88 (2012).

386. Grigoriev, I. V. *et al.* The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* **40.** doi:10.1093/nar/gkr947, 26–32 (Jan. 2012).

387. D'Souza, M., Larsen, N. & Overbeek, R. Searching for patterns in genomic data. *Trends in genetics : TIG* **13.** doi:10.1016/S0168-9525(97)01347-4, 497–8 (1997).

388. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215.** doi:10.1016/S0022-2836(05)80360-2, 403–410 (Oct. 1990).

389. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21.** doi:10.1093/bioinformatics/bti551, 3448–9 (2005).

390. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16.** doi:10.1016/S0168-9525(00)02024-2, 276–7 (2000).

391. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9.** doi:10.1371/journal.pbio.1001177, e1001177 (Oct. 2011).

392. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12.** doi:10.1371/journal.pbio.1001889, e1001889 (June 2014).

393. Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281.** doi:10.1126/science.281.5374.200, 200–207 (July 1998).

394. Scala, S., Carels, N., Falciatore, A., Chiusano, M. L. & Bowler, C. Genome properties of the diatom Phaeodactylum tricornutum. *Plant Physiol.* **129.** doi:10.1104/pp.010713, 993–1002 (July 2002).

395. Armbrust, E. V. *et al.* The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306.** doi:10.1126/science.1101156, 79–86 (Oct. 2004).

396. Lozano, J. C. *et al.* Efficient gene targeting and removal of foreign DNA by homologous recombination in the picoeukaryote Ostreococcus. *Plant J.* **78.** doi:10.1111/tpj.12530, 1073–1083 (June 2014).

397. Moulager, M., Corellou, F., Verge, V., Escande, M. L. & Bouget, F. Y. Integration of light signals by the retinoblastoma pathway in the control of S phase entry in the picophytoplanktonic cell Ostreococcus. *PLoS Genet.* **6.** doi:10.1371/journal.pgen.1000957, e1000957 (May 2010).

398. Dixon, L. E. *et al.* Light and circadian regulation of clock components aids flexible responses to environmental signals. *New Phytol.* **203.** doi:10.1111/nph.12853, 568–577 (July 2014).

399. Wagner, M. *et al.* Identification and characterization of an acyl-CoA:diacylglycerol acyltransferase 2 (DGAT2) gene from the microalga O. tauri. *Plant Physiol. Biochem.* **48.** doi:10.1016/j.plaphy.2010.03.008, 407–416 (June 2010).

400. Sorokina, O. *et al.* Microarray data can predict diurnal changes of starch content in the picoalga Ostreococcus. *BMC Syst Biol* **5.** doi:10.1186/1752-0509-5-36, 36 (2011).

401. Michely, S. *et al.* Evolution of codon usage in the smallest photosynthetic eukaryotes and their giant viruses. *Genome Biol Evol* **5.** doi:10.1093/gbe/evt053, 848–859 (2013).

402. Blanc-Mathieu, R., Sanchez-Ferandin, S., Eyre-Walker, A. & Piganeau, G. Organellar inheritance in the green lineage: insights from Ostreococcus tauri. *Genome Biol Evol* **5.** doi:10.1093/gbe/evt106, 1503–1511 (2013).

403. Kim, K. M., Park, J. H., Bhattacharya, D. & Yoon, H. S. Applications of next-generation sequencing to unravelling the evolutionary history of algae. *Int. J. Syst. Evol. Microbiol.* **64.** doi:10.1099/ijs.0.054221-0, 333–345 (Feb. 2014).

404. Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21.** doi:10.1101/gr.126599.111, 2224–2241 (Dec. 2011).

405. Narzisi, G. & Mishra, B. Comparing de novo genome assembly: the long and short of it. *PLoS ONE* **6.** doi:10.1371/journal.pone.0019175, e19175 (2011).

406. Aguilar, C. *et al.* Genetic changes during a laboratory adaptive evolution process that allowed fast growth in glucose to an Escherichia coli strain lacking the major glucose transport system. *BMC Genomics* **13.** doi:10.1186/1471-2164-13-385, 385 (2012).

407. Liu, H., Styles, C. A. & Fink, G. R. Saccharomyces cerevisiae S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics* **144,** 967–978 (Nov. 1996).

408. Bonhivers, M., Carbrey, J. M., Gould, S. J. & Agre, P. Aquaporins in Saccharomyces. Genetic and functional distinctions between laboratory and wild-type strains. *J. Biol. Chem.* **273.** doi:10.1074/jbc.273.42.27565, 27565–27572 (Oct. 1998).

409. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9.** doi:10.1371/journal.pgen.1003972, e1003972 (Nov. 2013).

410. Zhang, H., Meltzer, P. & Davis, S. RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* **14.** doi:10.1186/1471-2105-14-244, 244 (2013).

411. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11.** doi:10.1186/gb-2010-11-4-r41, R41 (2010).

412. Lam, H. M. *et al.* Glutamate-receptor genes in plants. *Nature* **396.** doi:10.1038/24066, 125–126 (Nov. 1998).

413. Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39.** doi:10.1093/nar/gkr425, e103 (Aug. 2011).

414. Ekblom, R., Smeds, L. & Ellegren, H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics* **15.** doi:10.1186/1471-2164-15-467, 467 (2014).

415. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95.** doi:10.1016/j.ygeno.2010.03.001, 315–327 (June 2010).

416. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40.** doi:10.1093/nar/gks001, e72 (May 2012).

417. Price, M. B., Jelesko, J. & Okumoto, S. Glutamate receptor homologs in plants: functions and evolutionary origins. *Front Plant Sci* **3.** doi:10.3389/fpls.2012.00235, 235 (2012).

418. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497.** doi:10.1038/nature12121, 127–131 (May 2013).

419. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp324, 1754–1760 (July 2009).

420. Abby, S. S., Touchon, M., De Jode, A., Grimsley, N. & Piganeau, G. Bacteria in Ostreococcus tauri cultures - friends, foes or hitchhikers? *Front Microbiol* **5.** doi:10.3389/fmicb.2014.00505, 505 (2014).

421. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5.** doi:10.1186/gb-2004-5-2-r12, R12 (2004).

422. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive misassembly. *Genome Biol.* **9.** doi:10.1186/gb-2008-9-3-r55, R55 (2008).

423. Philippe, N., Salson, M., Commes, T. & Rivals, E. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.* **14.** doi:10.1186/gb-2013-14-3-r30, R30 (2013).

424. Sonnhammer, E. L., von Heijne, G. & Krogh, A. *A hidden Markov model for predicting transmembrane helices in protein sequences* in. **6** (1998), 175–182.

425. Monnier, A. *et al.* Orchestrated transcription of biological processes in the marine picoeukaryote Ostreococcus exposed to light/dark cycles. *BMC Genomics* **11.** doi:10.1186/1471-2164-11-192, 192 (2010).

426. Foissac, S. & Schiex, T. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* **6.** doi:10.1186/1471-2105-6-25, 25 (2005).

427. Den Hartog, C. & Kuo, J. English. in *SEAGRASSES: BIOLOGY, ECOLOGY AND CONSERVATION* (eds Larkum, A. W. D., Orth, R. J. & Duarte, C. M.) doi:10.1007/978-1-4020-2983-7_1, 1–23 (Springer Netherlands, 2006). ISBN: 978-1-4020-2942-4. doi:10.1007/978-1-4020-2983-7_1.

428. Berry, J. A., Beerling, D. J. & Franks, P. J. Stomata: key players in the earth system, past and present. *Current Opinion in Plant Biology* **13.** doi:10.1016/j.pbi.2010.04.013, 232–239 (2010).

429. Aquino, R. S., Landeira-Fernandez, A. M., Valente, A. P., Andrade, L. R. & Mouro, P. A. S. Occurrence of sulfated galactans in marine angiosperms: evolutionary implications. *Glycobiology* **15.** doi:10.1093/glycob/cwh138, 11–20 (2005).

430. Larkum, A., Drew, E. & Ralph, P. English. in *SEAGRASSES: BIOLOGY, ECOLOGY AND CONSERVATION* (eds Larkum, A. W. D., Orth, R. J. & Duarte, C. M.) doi:10.1007/978-1-4020-2983-7_14, 323–345 (Springer Netherlands, 2006). ISBN: 978-1-4020-2942-4. doi:10.1007/978-1-4020-2983-7_14.

431. Franssen, S. U. *et al.* Transcriptomic resilience to global warming in the seagrass Zostera marina, a marine foundation species. *Proceedings of the National Academy of Sciences* **108.** doi:10.1073/pnas.1107680108, 19276–19281 (2011).

432. Mazzuca, S. *et al.* Establishing research strategies, methodologies and technologies to link genomics and proteomics to seagrass productivity, community metabolism, and ecosystem carbon fluxes. *Frontiers in Plant Science* **4.** doi:10.3389/fpls.2013.00038, 38– (2013).

433. Duarte, C. M. *et al.* Will the Oceans Help Feed Humanity? *BioScience* **59.** doi:10.1525/bio.2009.59.11.8, 967–976 (2009).

434. Les, D. H. & Tippery, N. P. English. in *Early Events in Monocot Evolution* (eds Wilkin, P. & Mayo, S. J.) doi:10.1017/CBO9781139002950.007, 118–164 (Cambridge University Press, 2013). doi:10.1017/CBO9781139002950.007.

435. Costanza, R. *et al.* The value of the world's ecosystem services and natural capital. *Nature* **387.** doi:10.1038/387253a0, 253–260 (1997).

436. Fourqurean, J. W. *et al.* Seagrass ecosystems as a globally significant carbon stock. *Nature Geosci* **5.** doi:10.1038/ngeo1477, 505–509 (2012).

437. Jaffe, D. B. *et al.* Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Research* **13.** doi:10.1101/gr.828403, 91–96 (2003).

438. Wang, W. *et al.* The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat Commun* **5.** doi:10.1038/ncomms4311, 3311 (2014).

439. Chavez Montes, R. A. *et al.* Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat Commun* **5.** doi:10.1038/ncomms4722, 3722 (2014).

440. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **369.** doi:10.1098/rstb.2013.0353. doi:10.1098/rstb.2013.0353 (2014).

441. Nauheimer, L., Metzler, D. & Renner, S. S. Global history of the ancient monocot family Araceae inferred with models accounting for past continental positions and previous ranges based on fossils. *New Phytologist* **195.** doi:10.1111/j.1469-8137.2012.04220.x, 938–950 (2012).

442. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Research.* doi:10.1101/gr.168997.113. doi:10.1101/gr.168997.113 (2014).

443. Golicz, A. A. *et al.* Genome-wide survey of the seagrass Zostera muelleri suggests modification of the ethylene signalling network. *Journal of Experimental Botany.* doi:10.1093/jxb/eru510. doi:10.1093/jxb/eru510 (2015).

444. Kirk, J. T. O. *Light and Photosynthesis in Aquatic Ecosystems* doi:10.1017/CBO9781139168212. ISBN: 9780521151757. doi:10.1017/CBO9781139168212 (Cambridge University Press, 2011).

445. Popper, Z. A. *et al.* Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. *Annual Review of Plant Biology* **62.** doi:10.1146/annurev-arplant-042110-103809, 567–590 (2011).

446. Touchette, B. W. Seagrass-salinity interactions: Physiological mechanisms used by submersed marine angiosperms for a life at sea. *Journal of Experimental Marine Biology and Ecology* **350.** doi:10.1016/j.jembe.2007.05.037, 194–215 (2007).

447. Michel, G., Tonon, T., Scornet, D., Cock, J. M. & Kloareg, B. The cell wall polysaccharide metabolism of the brown alga Ectocarpus siliculosus. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *New Phytologist* **188.** doi:10.1111/j.1469-8137.2010.03374.x, 82–97 (2010).

448. Colln, J. *et al.* Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proceedings of the National Academy of Sciences* **110.** doi:10.1073/pnas.1221259110, 5247–5252 (2013).

449. Hanson, S. R., Best, M. D. & Wong, C.-H. Sulfatases: Structure, Mechanism, Biological Activity, Inhibition, and Synthetic Utility. *Angewandte Chemie International Edition* **43.** doi:10.1002/anie.200300632, 5736–5763 (2004).

450. Furness, C. A. in *Early Events in Monocot Evolution* (eds Wilkin, P. & Mayo, S. J.) doi:10.1017/CBO9781139002950.005, 82–98 (Cambridge University Press, 2013). ISBN: 9781139002950.

451. Kuo, J. & den Hartog, C. English. in *SEAGRASSES: BIOLOGY, ECOLOGY AND CONSERVATION* (eds Larkum, A. W. D., Orth, R. J. & Duarte, C. M.) doi:10.1007/978-1-4020-2983-7$_3$, 51–87 (Springer Netherlands, 2006). ISBN: 978-1-4020-2942-4.

452. Cock, A. D. Flowering, pollination and fruiting in Zostera marina L. *Aquatic Botany* **9.** doi:10.1016/0304-3770(80)90023-6, 201–220 (1980).

453. Orth, R. J. *et al.* A Global Crisis for Seagrass Ecosystems. *BioScience* **56.** doi:10.1641/0006-3568(2006)56[987:AGCFSE]2.0.CO;2, 987–996 (2006).

454. Waycott, M. *et al.* Accelerating loss of seagrasses across the globe threatens coastal ecosystems. *Proceedings of the National Academy of Sciences* **106.** doi:10.1073/pnas.0905620106, 12377–12381 (2009).

455. Macreadie, P., Schliep, M., Rasheed, M., Chartrand, K. & Ralph, P. Molecular indicators of chronic seagrass stress: A new era in the management of seagrass ecosystems? *Ecological Indicators* **38.** doi:10.1016/j.ecolind.2013.11.017, 279–281 (2014).

456. Olsen, J. *et al.* Eelgrass Zostera marina populations in northern Norwegian fjords are genetically isolated and diverse. *Marine Ecology Progress Series* **486.** doi:10.3354/meps10373, 121–132 (2013).

457. Den Hartog, C., Hennen, J., Noten, T. & van Wijk, R. Chromosome numbers of the European seagrasses. *Plant Systematics and Evolution* **156.** doi:10.1007/BF00937201, 55–59 (1987).

458. Kuo, J. Chromosome numbers of the Australian Zosteraceae. *Plant Systematics and Evolution* **226.** doi:10.1007/s006060170063, 155–163 (2001).

459. Reusch, T. & Bostrm, C. Widespread genetic mosaicism in the marine angiosperm Zostera marina is correlated with clonal reproduction. *Evolutionary Ecology* **25.** doi:10.1007/s10682-010-9436-8, 899–913 (2011).

460. Doyle, J. J. Isolation of plant DNA from fresh tissue. *Focus* **12,** 13–15 (1990).

461. Maumus, F. & Quesneville, H. Deep Investigation of <italic>Arabidopsis thaliana</italic> Junk DNA Reveals a Continuum between Repetitive Elements and Genomic Dark Matter. *PLoS ONE* **9.** doi:10.1371/journal.pone.0094101, e94101 (Apr. 2014).

462. Quesneville, H. *et al.* Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLoS Comput Biol* **1.** doi:10.1371/journal.pcbi.0010022, e22 (July 2005).

463. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29.** doi:10.1038/nbt.1883, 644–652 (July 2011).

464. Gouzy, J., Carrere, S. & Schiex, T. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp024, 670–671 (2009).

465. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12.** doi:10.1186/1471-2105-12-323, 323 (2011).

466. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22.** doi:10.1093/bioinformatics/btl158, 1658–1659 (2006).

467. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btp120, 1105–1111 (2009).

468. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23.** doi:10.1093/bioinformatics/btl567, 257–258 (2007).

469. Zhang, L. *et al.* A Genome-Wide Characterization of MicroRNA Genes in Maize. *PLoS Genet* **5.** doi:10.1371/journal.pgen.1000716, e1000716 (Nov. 2009).

470. Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24.** doi:10.1093/bioinformatics/btn429, 2395–2396 (2008).

471. Meyers, B. C. *et al.* Criteria for Annotation of Plant MicroRNAs. *The Plant Cell* **20.** doi:10.1105/tpc.108.064311, 3186–3190 (2008).

472. Addo-Quaye, C., Miller, W. & Axtell, M. J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* **25.** doi:10.1093/bioinformatics/btn604, 130–131 (2009).

473. Wissler, L. *et al.* Dr. Zompo: an online data repository for Zostera marina and Posidonia oceanica ESTs. *Database* **2009.** doi:10.1093/database/bap009 (2009).

474. Van Bel, M. *et al.* Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. *Plant Physiology* **158.** doi:10.1104/pp.111.189514, 590–600 (2012).

475. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research* **18.** doi:10.1101/gr.081612.108, 1979–1990 (2008).

476. Stanke, M., Tzvetkova, A., Morgenstern, B., *et al.* AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7.** doi:10.1186/gb-2006-7-s1-s11, S11 (2006).

477. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **43.** doi:10.1093/nar/gku1243, D213–D221 (2015).

478. Consortium, T. U. UniProt: a hub for protein information. *Nucleic Acids Research* **43.** doi:10.1093/nar/gku989, D204–212 (2014).

479. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of Genome Duplications from Age Distributions Revisited. *Molecular Biology and Evolution* **30.** doi:10.1093/molbev/mss214, 177–190 (2013).

480. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11,** 725–736 (1994).

481. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24.** doi:10.1093/molbev/msm088, 1586–1591 (2007).

482. Fostier, J. *et al.* A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27.** doi:10.1093/bioinformatics/btr008, 749–756 (2011).

483. stlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38.** doi:10.1093/nar/gkp931, D196–D203 (2010).

484. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488.** doi:10.1038/nature11241, 213–217 (Aug. 2012).

485. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29.** doi:10.1093/molbev/mss075, 1969–1973 (2012).

486. GROUP, T. A. P. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161.** doi:10.1111/j.1095-8339.2009.00996.x, 105–121 (2009).

487. Gandolfo, M., Nixon, K. & Crepet, W. A new fossil flower from the Turonian of New Jersey: Dressiantha bicarpellata gen. et sp. nov. (Capparales). *American Journal of Botany* **85,** 964 (1998).

488. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **107.** doi:10.1073/pnas.0909766107, 18724–18728 (2010).

489. Crepet, W. & Nixon, K. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *American Journal of Botany* **85,** 1122 (1998).

490. Xi, Z. *et al.* Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences* **109.** doi:10.1073/pnas.1205818109, 17519–17524 (2012).

491. Doyle, J. A., Endress, P. K. & Upchurch, G. R. Early Cretaceous monocots: a phylogenetic evaluation. *Sbornk Nrodnho muzea v Praze* **B,** 59–87 (2008).

492. Iles, W. J. D., Smith, S. Y., Gandolfo, M. A. & Graham, S. W. Monocot fossils suitable for molecular dating analyses. *Botanical Journal of the Linnean Society* **178.** doi:10.1111/boj.12233, 346–374 (2015).

493. Janssen, T. & Bremer, K. The age of major monocot groups inferred from 800+ rbcL sequences. *Botanical Journal of the Linnean Society* **146.** doi:10.1111/j.1095-8339.2004.00345.x, 385–398 (2004).

494. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proceedings of the National Academy of Sciences* **107.** doi:10.1073/pnas.1001225107, 5897–5902 (2010).

495. Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytologist* **192.** doi:10.1111/j.1469-8137.2011.03794.x, 266–301 (2011).

496. Heled, J. & Drummond, A. J. Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation. *Systematic Biology* **61.** doi:10.1093/sysbio/syr087, 138–149 (2012).

497. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40.** doi:10.1093/nar/gkr1293, e49 (2012).

498. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13.** doi:10.1101/gr.1224503, 2178–2189 (2003).

499. Proost, S. *et al.* PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research* **43.** doi:10.1093/nar/gku986, D974–D981 (2015).

500. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) version 3.6* Seattle, 2005.

501. Maumus, F. & Quesneville, H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun* **5.** doi:10.1038/ncomms5104. doi:10.1038/ncomms5104 (2014).

502. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436.** doi:10.1038/nature03895, 793–800 (2005).

503. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463.** doi:10.1038/nature08747, 763–8 (2010).

504. Kikuchi, S. *et al.* Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* **339.** doi:10.1126/science.1229262, 571–574 (Feb. 2013).

505. Lohse, M., Drechsel, O., Kahlau, S. & Bock, R. OrganellarGenomeDRAW, a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41.** doi:10.1093/nar/gkt289, W575–W581 (2013).

506. Van den Berg, B. H., McCarthy, F. M., Lamont, S. J. & Burgess, S. C. Re-annotation is an essential step in systems biology modeling of functional genomics data. *PLoS One* **5.** doi:10.1371/journal.pone.0010642, e10642 (2010).

507. Scaife, M. A. *et al.* Establishing Chlamydomonas reinhardtii as an industrial biotechnology host. *The Plant Journal* **82.** doi:10.1111/tpj.12781, 532–546 (2015).

508. Misra, N., Panda, P. K., Parida, B. K. & Mishra, B. K. Phylogenomic study of lipid genes involved in microalgal biofuel production-candidate gene mining and metabolic pathway analyses. *Evol Bioinform Online* **8.** doi:10.4137/EBO.S10159, 545–64 (2012).

509. Krumholz, E. W., Yang, H., Weisenhorn, P., Henry, C. S. & Libourel, I. G. Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *Journal of Experimental Botany* **63.** doi:10.1093/jxb/err407, 2353–2362 (2011).

510. Petrie, J. R. *et al.* Metabolic engineering of omega-3 long-chain polyunsaturated fatty acids in plants using an acyl-CoA Delta6-desaturase with omega3-preference from the marine microalga Micromonas pusilla. *Metabolic engineering* **12.** doi:10.1016/j.ymben.2009.12.001, 233–40 (2010).

511. Tavares, S. *et al.* Metabolic Engineering of Saccharomyces cerevisiae for Production of Eicosapentaenoic Acid, Using a Novel ?5-Desaturase from Paramecium tetraurelia. *Applied and Environmental Microbiology* **77.** doi:10.1128/AEM.01935-10, 1854–1861 (2011).

512. Foresi, N. *et al.* Expression of the tetrahydrofolate-dependent nitric oxide synthase from the green alga Ostreococcus tauri increases tolerance to abiotic stresses and influences stomatal development in Arabidopsis. *The Plant Journal* **82.** doi:10.1111/tpj.12852, 806–821 (2015).

513. Blatt, A., Bauch, M. E., Prschke, Y. & Lohr, M. A lycopene -cyclase/lycopene e-cyclase/light-harvesting complex-fusion protein from the green alga Ostreococcus lucimarinus can be modified to produce a-carotene and -carotene at different ratios. *The Plant Journal* **82.** doi:10.1111/tpj.12826, 582–595 (2015).

514. Raffaele, S. *et al.* Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330.** doi:10.1126/science.1193070, 1540–3 (2010).

515. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog* **8.** doi:10.1371/journal.ppat.1002608, e1002608 (2012).

516. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* **14.** doi:10.1186/gb-2013-14-5-r51, R51 (2013).

517. Bellec, L. *et al.* Cophylogenetic interactions between marine viruses and eukaryotic picophytoplankton. *BMC Evolutionary Biology* **14.** doi:10.1186/1471-2148-14-59, 59 (2014).

518. Derelle, E. *et al.* Diversity of Viruses Infecting the Green Microalga Ostreococcus lucimarinus. *Journal of Virology* **89.** doi:10.1128/JVI.00246-15, 5812–5821 (2015).

519. Avrani, S., Wurtzel, O., Sharon, I., Sorek, R. & Lindell, D. Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature* **474.** doi:10.1038/nature10172, 604–8 (2011).

520. Meunier, J. & Duret, L. Recombination Drives the Evolution of GC-Content in the Human Genome. *Molecular Biology and Evolution* **21.** doi:10.1093/molbev/msh070, 984–990 (2004).

521. Bergero, R. & Charlesworth, D. The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol* **24.** doi:10.1016/j.tree.2008.09.010, 94–102 (2009).

522. Suda, S., Watanabe, M. M. & Inouye, I. EVIDENCE FOR SEXUAL REPRODUCTION IN THE PRIMITIVE GREEN ALGA NEPHROSELMIS OLIVACEA (PRASINOPHYCEAE)1. *Journal of Phycology* **25.** doi:10.1111/j.1529-8817.1989.tb00266.x, 596–600 (1989).

523. Soo Chan Lee, S., Ni, M., Li, W., Shertz, C. & Heitman, J. The evolution of sex: a perspective from the fungal kingdom. *Microbiol Mol Biol Rev* **74.** doi:10.1128/MMBR.00005-10, 298–340 (2010).

524. Ahmed, S. *et al.* A haploid system of sex determination in the brown alga Ectocarpus sp. *Curr Biol* **24.** doi:10.1016/j.cub.2014.07.042, 1945–57 (2014).

525. Bachtrog, D. *et al.* Sex determination: why so many ways of doing it? *PLoS Biol* **12.** doi:10.1371/journal.pbio.1001899, e1001899 (2014).

526. Hallmann, A. Evolution of reproductive development in the volvocine algae. *Sex Plant Reprod* **24.** doi:10.1007/s00497-010-0158-4, 97–112 (2011).

527. Bram, V. *Study of the introner elements found in the Micromonas genomes* master dissertation. 2010.

528. Dolezel, J. *et al.* Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomics* **12.** doi:10.1007/s10142-012-0293-0, 397–416 (2012).

529. Fink, G. R. Pseudogenes in yeast? *Cell* **49,** 5–6 (1987).

530. Sverdlov, A. V., Babenko, V. N., Rogozin, I. B. & Koonin, E. V. Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* **338.** doi:10.1016/j.gene.2004.05.027, 85–91 (2004).

531. Lin, K. & Zhang, D.-Y. The excess of 5' introns in eukaryotic genomes. *Nucleic Acids Research* **33.** doi:10.1093/nar/gki970, 6522–6527 (2005).

532. Simmons, M. P. *et al.* Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations. *Molecular Biology and Evolution* **32.** doi:10.1093/molbev/msv122, 2219–2235 (2015).

533. Chan, Y. A., Hieter, P. & Stirling, P. C. Mechanisms of genome instability induced by RNA-processing defects. *Trends Genet* **30.** doi:10.1016/j.tig.2014.03.005, 245–53 (2014).

534. Moore, M. J. & Sharp, P. A. Site-specific modification of pre-mRNA: the 2'-hydroxyl groups at the splice sites. *Science* **256,** 992–7 (1992).

535. Piégu, B., Bire, S., Arensburger, P. & Bigot, Y. A survey of transposable element classification systems A call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular Phylogenetics and Evolution* **86.** doi:10.1016/j.ympev.2015.03.009, 90–109 (2015).

536. Michel, F., Umesono, K. & Ozeki, H. Comparative and functional anatomy of group II catalytic introns–a review. *Gene* **82,** 5–30 (1989).

Shut up, Meg.
- Peter Griffin -