SEQUENCING AND ANALYSIS OF THE DIEL TRANSCRIPTOME OF *BOTRYOCOCCUS BRAUNII*

Submitted by Charlotte Cook to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Biological Sciences
In September 2014

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:

Charlotte Cook

“But I don’t want to go among mad people," Alice remarked.
"Oh, you can’t help that," said the Cat: "we’re all mad here.
I’m mad. You’re mad."
"How do you know I’m mad?" said Alice.
"You must be," said the Cat, "or you wouldn’t have come here.”
— Lewis Carroll, Alice in Wonderland

ACKNOWLEDGEMENTS

This study was funded by Biotechnology and Biological Sciences Research Council (BBSRC), additional funding was provided by Plymouth Marine Laboratories (PML) as CASE partner.

Firstly I would like to thank my supervisors; Dr John Love and Dr Christine Sambles of the University of Exeter and Dr Mike Allen of PML. John, thank you for your enthusiasm and providing me with this opportunity and others that many PhD students do not have the benefit of experiencing. Christine, I am particularly indebted to you for your invaluable support, understanding and technical training, especially that given prior to your official recognition as my supervisor. Mike, thank you for providing me with encouragement, advice and fresh perspectives on my research.

I would like to thank those who contributed to the work in this project; namely Dr Karen Moore and Audrey Farbos of the University of Exeter Sequencing Service, who advised on the experimental design and provided training on library preparation and sequencing protocols. Dr Mike Page, thank you for your advice, training, comedy and tunes in the lab.

I am grateful to the Mezzanine Lab team for providing me with a stimulating and bizarrely entertaining work environment. If I hadn't laughed, I might have cried or tried to escape the ward.

Completing a PhD is an excellent indicator of the great friends you have. I am grateful to Olly, for being there at the outset of this venture. Charlotte, Donna, Jen, Drs Mike and Rhiannon Page and the ever- forgiving Gatsby the Golden Retriever: thank you all for not giving up on me whilst I went off- radar and for providing a warm welcome home and enough wine when I was in the proximity.

I can't thank my parents enough for their love, understanding and stoical support, in all guises that it was received and intuitively given in exactly the measure I needed.

ABSTRACT

Microalgae are widely viewed as a potential source of renewable biofuels. Microalgae are highly productive and can be cultured in recycled water on margial or non-agricultural land. Despite their advantages, the industrial scale deployment of microalgae faces numerous challenges including relatively little knowledge of the algae themselves and the comparatively expensive infrastructures required for culture.

The green microalga*, Botryococcus braunii* is particularly interesting because it synthesizes long-chain (C30- C40) hydrocarbons that can be converted to liquid fuel by hydrogenation and catalytic cracking. Moreover, *B. braunii* is the major fossil present in the Ordovician oil shales and kerogen deposits. Although studied since the 1970s, very little is known regarding critical aspects of *B. braunii*, notably its molecular biology. In higher plants molecular clocks have been well defined and transcript profiling has revealed a sophisticated network of circadian scheduling of metabolic processes. Characterization of temporal controls over hydrocarbon synthesis is therefore of importance to optimization of biofuel production from *B. braunii*.

In this project *B. braunii* (Race B, strain Guadeloupe) were cultured in a 12-hour photoperiod and either maintained in that regime or transferred to constant light. Algae were sampled every 4 hours, during a 28-hour time-course and mRNA extracted. mRNA was reverse-transcribed to cDNA and sequenced using a paired-end protocol on an Illumina HiSeq 2000 platform. Over 2 billion sequence reads of 100 bp were generated and assembled *de novo*, into a complete transcriptome for *B. braunii*. The transcriptome was comprehensively annotated using global and targeted protocols and differential expression and co-expression analyses were performed.

Metabolic pathway analysis confirmed the presence, and photoperiodic regulation of the MEP/DOXP Terpenoid Backbone synthesis pathway. Targeted annotation and expression analysis revealed two predicted *B. braunii* circadian clock components, which were incorporated into a *B. braunii* circadian clock model. In non-hierarchical cluster analysis, contigs of the *B. braunii* transcriptome clustered under four distinct patterns of diel expression. Networks of co- and anti-expressed contigs were elucidated by hierarchical clustering. These results demonstrate the exquisite control over metabolism in *B. braunii.* Such knowledge is essential for the industrial applications of *B. braunii*, either directly or through the engineering of selected *B. braunii* genes or molecular pathways into alternative chassis.

CONTENTS

ACKNOWLEDGEMENTS **3**

ABSTRACT **4**

ABBREVIATIONS **12**

CHAPTER 1: INTRODUCTION **15**
1.1 Global fuel supply **15**
1.2 Advanced biofuels from microalgae **16**
1.3 Microalgal culture systems **17**
1.4 Botryococcus braunii **21**
1.4.1 Significance 21
1.4.2 Physiology 21
1.4.3 Hydrocarbon production 25
1.4.4 Molecular pathways of hydrocarbon synthesis 26
1.5 Circadian clocks **27**
1.5.1 Importance of circadian control 27
1.5.2 Circadian control of metabolism 28
1.5.3 Circadian clock architecture in photosynthetic organisms 29
1.6 Identification of clock components **34**
1.7 DNA sequencing technologies **35**
1.7.1 Chemical and chain termination sequencing 35
1.7.2 PCR-based sequencing 36
1.7.3 Illumina sequencing 37
1.8 RNA-Seq **39**
1.8.1 The application of RNA-Seq to circadian biology and metabolism studies 39
1.8.2 De novo transcriptome assembly of B. braunii 40
1.9 Project Plan **40**

CHAPTER 2: GENERAL MATERIALS AND METHODS **44**
2.1 Media preparation **44**
2.1.1 Modified Chu 13 44
2.1.2 Luria Bertani broth agar preparation 44
2.1.3 Yeast Mould agar preparation 45
2.1.4 mRNA buffers 45
2.2 Algal culture conditions **45**
2.3 Monitoring potential contamination of algal cultures **45**
2.4 Absorbance of algal culture **46**
2.5 Algal cell harvest **46**
2.6 Quantification of dry biomass **46**
2.7 Quantification of chlorophyll **47**
2.8 Total RNA extraction from Botryococcus braunii cells **47**
2.9 Addition of external RNA controls **49**
2.10 mRNA purification **49**
2.11 Quantification and analysis of nucleic acids **49**
2.11.1 Nanodrop analysis 49
2.11.2 Agilent Bioanalyzer analysis 50
2.11.2.1 RNA analysis *50*
2.11.2.2 DNA analysis *50*
2.12 Construction of cDNA libraries and sequencing **51**

CHAPTER 3: FROM *BOTRYOCOCCUS BRAUNII* CULTURES TO ANNOTATED TRANSCRIPTOME **53**
3.1 Introduction **53**
3.2 Materials and Methods **57**
3.2.1 Harvest and preparation of samples 57
3.2.2 Construction of cDNA libraries and sequencing 60
3.2.3 Pre-processing of raw reads and sequence assembly 60
3.2.4 Functional annotation 65
3.2.4.1 BLAST sequence homology *65*
3.2.4.2 Gene Ontology, KEGG and EC features - Annot8r *66*
3.2.4.3 Protein family assignment - Pfam search *67*
3.2.5 Removal of contaminating sequences 67
3.3 Results **68**
3.3.1 Preparation of samples, construction of cDNA libraries and sequencing 68
3.3.2 Pre-processing of raw reads and sequence assembly 71
3.3.3 Functional annotation of assembled *B. braunii* transcriptome 79
3.3.3.1 BLAST search of NCBI Non-Redundant Protein Database *79*
3.3.3.2 Gene Ontology annotation *81*
3.3.3.3 KEGG pathway analysis *89*
3.3.3.4 Assignment of protein domain families using Pfam scan *95*
3.4 Discussion **97**
3.4.1 Pre-processing of raw reads and sequence assembly 97
3.4.2 Functional annotation 99
3.4.2.1 BLAST search of NCBI Non-Redundant Protein Database *99*
3.4.2.2 Gene Ontology annotation *100*
3.4.2.3 KEGG pathway analysis *101*
3.4.2.4 Assignment of protein domain families using Pfam scan *102*
3.4.3 Summary 103
Chapter 3 Summary **104**

CHAPTER 4: VALIDATING THE *BOTRYOCOCCUS BRAUNII* TRANSCRIPTOME **105**
4.1 Introduction **105**
4.2 Materials and Methods **111**
4.2.1 Identification of clock protein homologs by sequence 111
4.2.2 Identification of clock protein homologs by functional domain 113
4.2.2.1 Domain and motif identification *113*
4.2.2.2 MUSCLE alignment of Botryococcus braunii putative clock components to known clock components *113*
4.2.2.3 HMM alignment of Botryococcus braunii predicted clock components to model clock components *113*
4.2.2.4 Keyword search of HMM Botryococcus braunii database *113*
4.2.2.5 Clock component HMM generation and transcriptome HMM scan *113*
4.3 Results **115**
4.3.1 Identification of clock protein homologs by sequence 115
4.3.2 Identification of clock protein homologs by functional domain 118
4.3.2.1 Domain and motif identification *118*
4.3.2.2 Domain architecture comparison with model clock components *120*
4.3.2.3 Custom HMM generation and scan of Botryococcus braunii ORFs *135*
4.3.2.4 Keyword domain search of Botryococcus braunii ORFs *139*
4.3.2.5 The Botryococcus braunii clock model *140*
4.4 Discussion **142**
4.4.1 Identification of clock protein homologs by sequence 142
4.4.2 Identification of clock protein homologs by functional domain 143
4.4.2.1 Domain and motif identification *143*
4.4.2.2 Domain architecture comparison with model clock components *143*
4.4.2.3 Custom HMM generation and scan of Botryococcus braunii ORFs *146*
4.4.2.4 Keyword domain search of Botryococcus braunii ORFs *146*

4.4.3 The Botryococcus braunii clock model 147
Chapter 4 Summary 149

CHAPTER 5: ANALYSIS OF TIME- DEPENDENT DIFFERENTIAL EXPRESSION IN THE *BOTRYOCOCCUS BRAUNII* TRANSCRIPTOME 150
5.1 Introduction 150
5.2 Materials and Methods 153
5.2.1 Alignment of timepoint reads to reference transcriptome assembly 153
5.2.2 Generation of count data using HTSeq 153
5.2.3 Differential expression analysis by DESeq2 154
5.2.4 Removal of outliers 155
5.2.4.1 Candida glabrata contamination 155
5.2.4.2 Screening for anomalous patterns 155
5.2.4.3 Removal of samples with low read alignment number 156
5.2.5 Targeted expression analysis of genes of interest 156
5.2.5.1 DE of predicted clock components and terpenoid pathway genes 156
5.2.5.2 Expression patterns of predicted clock genes 157
5.2.6 Cluster analysis of expression 157
5.2.6.1 Scree plots and K-means 157
5.2.6.2 Pearson correlation 158
5.3 Results 159
5.3.1 Alignment of timepoint reads to reference transcriptome assembly 159
5.3.2 Outliers 159
5.3.3 DE between conditions 159
5.3.4 Temporal DE 162
5.3.4.1 Constitutively expressed genes and pathways 163
5.3.4.2 Genes and pathways with DE 169
5.3.4.3 Genes with light dependent differential expression 176
5.3.5 Expression of predicted *Botryococcus braunii* clock components 184
5.3.5.1 Expression profile of predicted Botryococcus braunii PRR1- like transcript 184
5.3.5.2 Expression of predicted Botryococcus braunii CCA1-like transcript 186
5.3.5.3 Comparative analysis of expression patterns 188
5.3.6 Co-expression networks and cluster analysis 192
5.3.6.1 K-means clustering 185
5.3.6.2 Pearson correlation of expression profiles 198
5.4 Discussion 200
5.4.1 Contextualisation of bioinformatics methods 200
5.4.1.1 Read alignment using Bowtie 200
5.4.1.2 Read quantification using HTSeq-count 200
5.4.1.3 Statistical analysis of differential expression 201
5.4.1.4 RNA-Seq contamination and dataset integrity 203
5.4.2 Differentially expressed transcripts 203
5.4.3 Expression of sequences involved in terpene biosynthesis 204
5.4.4 Expression profiles of the predicted circadian clock sequences 206
5.4.5 Co- expression analysis of the differentially expressed transcripts 208
5.5 Summary 210

DISCUSSION 211
6.1 The potential of Next Generation Sequencing for investigating microalgal physiology 211
6.2 Sequencing, annotation and analysis of the Botryococcus braunii transcriptome 212
6.3 Identification and characterisation of circadian clock components of the Botryococcus braunii transcriptome 214
6.4 Mapping and characterisation of the Botryococcus braunii triterpenoid synthesis pathway 216
6.5 Regulatory networks of gene expression in the Botryococcus braunii transcriptome 217

CONCLUSION **220**

APPENDIX **223**

BIBLIOGRAPHY **250**

FIGURES

CHAPTER 1: Introduction
Figure 1 Tubular photobioreactors 19
Figure 2 Tubular-flat-plate hybrid photobioreactor 20
Figure 3 Botryococcus braunii "Guadeloupe" at 100 x magnification 23
Figure 4 Confocal image of Botryococcus braunii "Guadeloupe" 24
Figure 5 Arabidopsis thaliana circadian clock model 31
Figure 6 Ostreococcus tauri circadian clock model 33
Figure 7 Diagramatic representation of Illumina sequencing 38
Figure 8 Schematic representation of sample processing 42
Figure 9 Workflow showing experimental and bioinformatic procedures 43
CHAPTER 3: From *Botryococcus braunii* cultures to annotated transcriptome
Figure 1 Workflow showing experimental and bioinformatic procedures 56
Figure 2 Sample timeline 58
Figure 3 Schematic of sample processing 59
Figure 4 Illumina quality scoring 62
Figure 5 Generation of contigs by De Bruijn graphs 64
Table 1 RNA quality and concentration by sample 69
Figure 6 Bioanalyzer trace images 70
Table 2 Sequence reads by lane 72
Figure 7 Raw read FastQC quality checks 74
Figure 8 Processed read FastQC quality checks 75
Table 3 Assembly statistics 77
Figure 9 Contig length distribution 78
Figure 10 Taxonomic distribution of BLAST annotations 80
Table 5 Assignment of read counts to GO categories 82
Figure 11 Distribution of Biological Process GOs 84
Figure 12 Distribution of Molecular Function GOs 86
Figure 13 Distribution of Cellular Component GOs 88
Figure 14 KEGG pathways map of *B. braunii* transcriptome 91
Table 5 Top 20 KEGG Pathways 92
Figure 15 KEGG terpenoid backbone pathway 93
Figure 16 KEGG squalene synthesis pathway 94
Figure 17 Top 20 Pfam domains in *B. braunii* transcriptome 96
CHAPTER 4: Validating the *Botryococcus braunii* transcriptome
Figure 1 Workflow of experimental and bioinformatic methods 108
Figure 2 Higher plant model 109
Figure 3 *Ostreococcus tauri* clock model 110

Table 1 Clock protein BLAST queries 112
Table 2 BLAST protein sequence homology results 117
Table 3 Model clock component and *B. braunii* sequence homolog domains 119
Figure 4 PRR1/ TOC1 domains 122
Figure 5 PRR3 domains 124
Figure 6 PRR5 domains 126
Figure 7 PRR7 domains 128
Figure 8 PRR9 domains 130
Figure 9 CCA1/LHY- like component domains 132
Figure 10 CHLAMY1 subunit C1 domains 133
Table 4 Top 30 pseudo response regulator HMM ORFs 135
Table 5 Top 30 myb DNA binding HMM ORFs 137
Figure 11 *B. braunii* clock model 140

CHAPTER 5: Analysis of time- dependent Differential Expression in the *Botryococcus braunii* transcriptome

Figure 1 Workflow for experimental and bioinformatic methods 152
Table 1 Subsets of transcripts 157
Figure 2 Expression dispersion of transcripts in LL and LD 161
Figure 3 KEGG overview Metabolic pathways map of *B.braunii* transcripts with no temporal DE 165
Figure 4 KEGG terpenoid backdone pathway with B.braunii transcripts with no time dependent DE mapped 167
Figure 5 KEGG Sesquiterpenoid reference pathway with non-temporally differentially expressed *B. braunii* transcripts mapped 168
Figure 6 KEGG Metabolic pathways overview mapped with *B. braunii* transcripts with DE 170
Figure 7 KEGG terpenoid backbone biosynthesis pathway mapped with temporally differentially expressed *B. braunii* transcripts 172
Figure 8 KEGG sesquiterpenoid pathway mapped with transcripts with DE in both LD and LL conditions 173
Figure 9 KEGG Plant circadian clock reference pathway mapped with *B. braunii* DE transcripts 175
Figure 10 KEGG Metabolic pathways overview of photoperiodic *B. braunii* transcripts 178
Figure 11 KEGG Terpenoid backbone biosynthesis pathway mapped with photoperiodic *B. braunii* transcripts 180
Figure 12 KEGG Sesquiterpenoid pathway mapped with photoperiod *B. braunii* transcripts 181
Figure 13 KEGG Circadian clock (plant) pathway mapped with photoperiodic *B. braunii* transcripts 183
Figure 14 mRNA expression of predicted *B. braunii* TOC1 185
Figure 15 mRNA expression of predicted *B. braunii* LHY/CCA1 187
Figure 16 Expression patterns of TOC1 189

Figure 17 Expression patterns of CCA1 191
Figure 18 *K*-means clustering of transcripts under LD conditions 194
Figure 19 *K*-means clustering of transcripts under LL conditions 197
Figure 20 Clustergram of transcript expression 199

ABBREVIATIONS

A	adenine
aa	amino acid
ABC	ATP binding cassette
ARR	authentic response regulator
ATP	adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	base pairs
C	cytosine
CCA1	circadian clock associated
cDNA	complementary deoxyribose- nucleic acid
CK1	casein kinase 1
CO_2	carbon dioxide
contig	consensus sequence (of nucleic acids) of a set of overlapping sequencing reads
CPU	central power unit
DE	differential expression
DNA	deoxyribose- nucleic acid
E value	expect value
EC	Enzyme Commission number
ECM	extracellular matrix
EE	evening element
ELF	early flowering
EROI	energy return on investment
EST	expressed sequence tag
G	guanine
GC	guanine- cytosine
GHG	green house gas
GO	Gene Ontology
HK	histidine kinase
HMM	Hidden Markov Model
IRR	internal rate of return
k-mer	a sequence of nucleic acids where k is any integer of choice

KEGG	Kyoto Encyclopedia of Genes and Genomes
KO ID	KEGG Orthology identifier
LCA	lowest common ancestor
LHY	late elongated hypotyl
LUX	LUX arrythmo
MA	M (log ratio) and A (mean average)
MAPK	mitogen activated protein kinase
MEP/ DOXP	methylerythritol 4-phosphate/deoxyxylulose phosphate pathway
mRNA	messenger ribose- nucleic acid
MSV	Multiple Segment Viterbi
MYA	million years ago
NGS	Next generation sequencing
nt	nucleotide
OH	hydroxyl group
Oligo(dT)	oligo- deoxythymine
ORF	open reading frame
padj	adjusted p value
PBRs	photobioreactors
Pkinase	protein kinase
Pkinase_TYR	protein tyrosine kinase
PP	posterior probability
PRR	pseudo- response- regulator
PSPP	pre-squalene diphospate
QPHRED	Phred quality score
RIN	RNA Integrity Number
RNA	ribose- nucleic acid
RNA-seq	RNA- sequencing
RR	response regulator
RRR	response regulator receiver
SAM	sequence alignment/ map
SBS	sequencing by synthesis
T	thymine
TAGs	triacylglycerols
tRNA	transport RNA
TTFL	transcriptional- translational feedback loop

CHAPTER 1

INTRODUCTION

1.1 Global fuel supply

Mobility of people and goods is essential to the current and future prosperity of the World. A secure supply of fossil fuel for shipping, aviation and private vehicles underpins that mobility. However, fossil fuels are increasingly rare and difficult to extract. Moreover burning fossil fuels releases into the atmosphere CO_2, which has been linked to anthropogenic climate change. Biofuel, *i.e.* fuels that are derived from living biomass rather than fossilised biomass, are widely accepted as a realistic supplement to fossil hydrocarbons. In the long term, it is anticipated that biofuels will eventually replace fossil hydrocarbons and provide the opportunity to diversify income and fuel supply sources, increase the security of energy supplies and sustain or enhance employment in rural areas (Mata *et al*., 2010).

First generation biofuels, ethanol or “biodiesels”, are primarily produced from crop biomass and seed oils. In retail fuel the proportion of first generation biofuels is mandated to at least 10% in the EU, the USA and Canada, with similar targets being enacted in China and India. However, the production of first generation biofuel is insufficient to meet global demand and competes directly with the production of food (Scarlet *et al*., 2008). This competition, known as the food versus fuel problem skews commodities markets worldwide and decreases food security most notably in developing countries, *i.e.* those who need it most. Moreover, the pressure for land use change towards an increase in cultivated fields may lead to the loss of biodiversity and important ecological areas (Gallagher *et al*., 2008). Second generation biofuels, derived from the fermentation of pre-treated, lignocellulosic biomass from plant cell walls, have the potential to reduce the food-fuel dilemma, but despite pilot production, are not yet commercialized at scale (Lehto *et al*., 2014). In some cases, first generation biofuels have a greater total environmental impact than fossil fuels. For example, deforestation to clear land for ethanol production from sugar cane releases greenhouse gases in a quantity large enough to negate the benefits of the biofuel (Sheehan *et al*., 2003). However, such cases are not the norm; of 26 biofuels studied, 21 were shown to decrease greenhouse gas emissions by 30% or more, although 12 had a greater

overall environmental impact than fossil fuel production when compared with gasoline production and use (Greenwell *et al*., 2010).

1.2 Advanced biofuels from microalgae

The large-scale culture of microalgae has been promoted as a potential opportunity to meet biofuels targets whilst reducing land competition with other arable crops. Microalgae have a potentially higher growth rate and productivity compared with terrestrial feedstocks, for example for 30 % w/w of oil content in algal biomass 39 to 132 times less land area is required for cultivation than for rapeseed or soybean (Chisti, 2007). Additionally, various fuels can be produced from microalgae including methane, hydrogen, biodiesel and ethanol. Algal biodiesel contains fewer particulates and emits less CO, sulphur oxides and hydrocarbons whilst performing as well as conventional diesel (Delucchi, 2003; Gouveia & Oliveira, 2008).

The photosynthetic and metabolic properties of microalgae are now broadly acknowledged as valuable to the biofuels production industry (Chisti, 2007; Donohue and Cogdell, 2006) as well as having various other applications. Neutral lipids, predominantly triacylglycerols (TAGs), produced in large quantities by microalgae may provide an alternative to conventional oils in the food industry, e.g. *Botryococcus sudeticus* produces large quantities of TAGs with a structure similar to olive oil (Vazquez-Duhalt & Greppin, 1987). Strains of algae are already in commercial use as human nutritional supplements because of their high protein and other nutritional benefits (Spolaore *et al*., 2006), for example health drinks are made from *Chlorella* (Becker, 1994). Functional peptides with applications in healthcare have been isolated from microalgae, representing an alternative to costly production of medication sourced from animal and plant protein (Sheih *et al*., 2009). *D. salina* and *Haematococcus pluviella* are both cultured commercially and experimentally respectively, at a large scale for carotenoid production (Ben-Amotz and Avron, 2014) with *Chlorella* sp., *Scenedesmus almeriensis* and *Muriellopsis* sp also demonstrating potential in this industry (Blanco *et al*., 2007; Del Campo *et al*., 2001; Matsukawa *et al*., 2000). Carotenoids are considered effective agents against certain human diseases (Krinsky *et al*., 2003). *Spirulina* extracts are used to produce blue colouring for cosmetic applications (Raja *et al*., 2008). After oil has been harvested the microalgal biomass can be processed into animal feed in agriculture or aquaculture (Neori, 2010), organic fertilizer, burned for energy generation or converted into ethanol or methane (Spolaore *et al*., 2006). These biotechnological applications are made further more attractive via the possibility of sequestration of CO_2 and other emission types from flue gas by algal

bio-fixation (Chiu *et al*., 2011) thereby reducing industrially derived Green House Gas (GHG) emissions and pollution, whilst producing biodiesel (Directive 2008/0016/COD of the European Parliament and of the council on the promotion of use of renewable energy sources 2008). Treatment of domestic and industrial wastewater; coupled with biomass production (with direct or indirect uses in energy production) offers an elegant and profitable biotechnological solution for clean usable water (Cabanelas *et al*., 2013; Lichtenthaler, 2014).

From a commercial perspective, biofuels are low-value, high-volume products and consequently the high cost of production is the main obstacle to production at a industrial scale. At a crude oil price of US$200 bbl^{-1} (double the actual cost at the time of writing), an alga with an oil content of 40% must cost less than US$0.45 kg^{-1} to produce (Borowitzka, 1999)- this is considerably less than the actual cost of commercial production (Borowitzka, 2013). However assuming that oil is extracted from microalgae grown over 500 ha, the co-production of high value products and the sale of remaining biomass as feedstock using technologies available today, microalgal-derived biofuel production is modelled to be economically viable according to a 30 year internal rate of return (IRR) of 15% (Stephens *et al*., 2010). These models, however, do not account for variation in commodity, fossil fuel or land price, but assume a long-term historical trajectory.

1.3 Microalgal culture systems

Different systems exist for algal culture at commercial scales, including central pivot ponds, raceways or photobioreactors (PBRs). For ponds or raceways dimensions range from 0.5 Ha to 200 Ha depending on location pond depths range from 0.2- 0.3 m to allow sunlight to penetrate through the water column. In central pivot ponds, a centrally pivoted paddle agitates the water to mix the algae. Raceway ponds operate by circulating algae, water and nutrients around a channel using a baffle to provide flow and removing algae containing water at one end by centrifugation or flocculation (Campo *et al*., 2007). A source of waste CO_2 may be bubbled in to the pond. Typical coal-fired power plant flue gas contains 13% CO_2, thus coupling an algal farm with a coal power plant provides a worthy solution to recycle CO_2 into a useful biomass crop.

Limitations to the commercial growth of microalgae in open ponds are reduction in photosynthetic efficiency because of poor mixing and long light path, high evaporative and CO_2 losses to the atmosphere, the large land area required, contamination and the expense of harvesting a low biomass concentration (Mata *et al*., 2010; Mirón *et al*., 2003). Despite these concerns, open ponds and raceways are

anticipated to be the most appropriate cultivation options for commercial purposes owing to their simplicity and low costs (Sheehan *et al*., 2003). Optimization of conditions in raceways and ponds is required for those microalgal species proposed for oil production (Rodolfi *et al*., 2009).

Surface area to volume ratio of microalgal culture in PBRs is maximized and evaporative losses, exposure to contaminants and CO_2 diffusion into the atmosphere are minimized by the closed system. Henceforth, growth of microalgae in PBRs can have a greater photosynthetic efficiency than in open systems (Arbib *et al*., 2013). Common types of PBR include tubular or flat plate. In tubular PBRs, algae are circulated by either a mechanical pump or an airlift system through tubes of glass, Plexiglass, Poly Vinyl C or polyethylene (Figure 1). Airlift systems result in reduced shear stress of the culture and can provide the means of gas exchange in the system (Mirón *et al*., 2003). Alternatively, algae may be grown in flat plates where solar light is laminated and there is a large illumination surface area yielding high productivity per unit of ground area. Tubular- flat-plate hybrid systems are also available, whereby a flat-plate system is provided with dense algal inoculum by a tubular PBR (Figure 2). However, biofouling, over-heating, accumulation of O_2, difficulty in the logistics of scale-up and the high cost of production and culture maintenance, which may be an order of magnitude higher than that in open systems, make PBRs an unlikely medium for microalgal culture on an industrial scale (Mirón *et al*., 2003; Ugwu *et al*., 2008).

Open ponds have been coupled with PBRs experimentally, with the goal of reducing time exposed to adverse events in ponds by providing sufficient clean inoculants (Rodolfi *et al*., 2009).

Figure 1 Tubular photobioreactors
Strains of microalgae growing in vertical tubular PBRs. Image courtesy of Dr Michael Allen (Plymouth Marine Laboratory).

Figure 2 Tubular-flat-plate hybrid photobioreactor
Microalgae growing to a dense innoculum within a vertical tubular PBR and feeding into a flat-plate PBR. Image courtesy of Dr Michael Allen (Plymouth Marine Laboraties).

Factors that pose a challenge to the commercialization of algal biofuels are mainly those that affect energy return on investment (EROI). Integrated bio-refineries provide an opportunity to maximize EROI, where algae are cultured and utilized for a purpose, preferably high value products of the kind outlined above, and biofuel becomes a by-product of this. Margins are in any case very tight, so an increase in process efficiency may have a disproportionate effect on the economic feasibility of microalgal biofuels production. Increases in process efficiency may be gained by modifying factors external to the microalgal culture, however the biological traits of the organisms themselves may also be capitalized upon, subject to improvements.

1.4 *Botryococcus braunii*

1.4.1 Significance

Botryococcus braunii is a colonial green microalga and is of interest because of prolific hydrocarbon production. Furthermore, *B. braunii* is implicated as one of the primary sources of present day fossil fuel reserves. The fossil history of *Botryococcus* type algae is long and records the genus from the Carboniferous period. *B. braunii* appears as the primary organism contributing to some of the richest oil shales in the world; torbanite (Temperley, 1936; Tyson, 1995b), the Eocene Green River shales in Colorado and Utah and the Ordovician Kukersite of Estonia (Glikson *et al*., 1989). Botryococcane, a hydrogenated botryococcene, is reported to comprise 1.4 % of Sumatran crude oil. Botryococcenes are acyclic isoprenoid compounds of the general formula C_nH_{2n-10} comprised of dimethylated carbon units (Cox *et al*., 1973; Metzger *et al*., 1988). There are no other reported occurrences of a single fossil marker in crude oil near the 1 % level (Moldowan & Seifert, 1980); hence the significance of *B. braunii* to the fossil oil industry.

1.4.2 Physiology

There are many strains of *B. braunii*, grouped into races A, B or L. Recently a fourth race, S, that is similar to the L race has been tentatively proposed (Kawachi *et al*., 2012). It has been suggested that strains of *B. braunii* belong to different species however there has been no definitive evidence to confirm this, in part due to morphological changes induced by different culture conditions of *B. braunii* (Volkman, 2014). Until recently members of the B race were classified under Chlorophycaeae, which is one of three monophyletic groups of the Chlorophyta (Sawayama *et al*., 1995). However subsequent phylogenetic analysis contradicted previous analyses and *B.*

braunii B race strains as well as some A and L, have been reclassified on the NCBI taxonomic database as belonging to the Trebouxiophyceae (Senousy *et al*., 2004).

B. braunii forms colonies of pyriform cells of 13 x 7-9 μm, embedded in an extracellular matrix (ECM) of polymerised and liquid hydrocarbons, which is surrounded by an outer retaining wall (Figure 3) (Metzger *et al*., 1988). In race B strains, the ECM is primarily composed of polyacetal hydrocarbons, which are covalently cross-linked with tetramethylsqualene diols and botryococcenes, which fill the interstitial space within colonies. The retaining wall is closely associated with the apical wall surface of individual cells. A polysaccharidic fibrillar sheath extends into the medium from the outside surface of the retaining wall. Each cell has an apical Golgi body and a cortical, fenestrated endoplasmic reticulum that is thought to participate in the secretion of the retaining wall and fibrillar sheath components. In the non- apical domain of B race cells, lipid inclusion bodies (Figure 4) containing botryococcenes are present and associate with the ER, nuclear and chloroplast envelopes. The non-apical ER is in close contact with the cell membrane from where it is proposed that botryococcenes are secreted via a continuous process, generating a layer of hydrocarbon emanating into the ECM from each cell (Weiss *et al*., 2012).

Retention of colonies within a high lipid content matrix allows them to float and therefore increase exposure to sunlight, maximising photosynthetic potential (Banerjee *et al*., 2002). Other drivers for prolific hydrocarbon production in *B. braunii* remain undefined. During stationary phase colonies turn from green to orange- red in colour as caretenoids accumulate (Metzger *et al*., 1988).

B. braunii is globally widespread with cultures isolated from Australia, Bolivia, Thailand, Ivory Coast and Israel amongst others (Grice *et al*., 1998; Metzger & Casadevall, 1987; Metzger *et al*., 1988; 1990). *B. braunii* occupies diverse aquatic environments, appearing in freshwater lakes, marine fjords and brackish reservoirs and ponds (Fuhrmann *et al*., 2003; Grice *et al*., 1998; Huang *et al*., 1999; Zhang *et al*., 2007). Despite its slow growth, *B. braunii* can form enormous blooms, such as those that have occurred in the Darwin river reservoir, Australia (Glikson *et al*., 1989; Tyson, 1995a; Wake & Hillen, 1980; 1981).

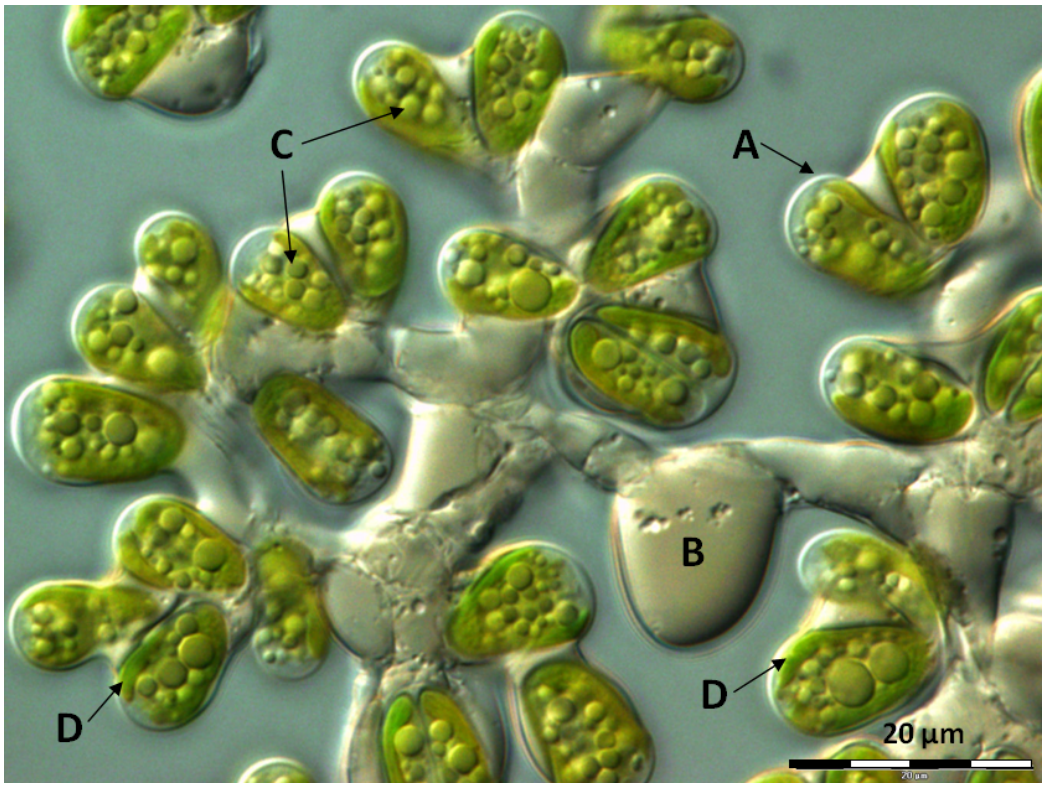


Figure 3 *Botryococcus braunii* "Guadeloupe" at 100 x magnification
Differential Interference Contrast (DIC) microscopy image of a partial *B. braunii* colony where the letters depict the following: (A) individual cell, (B) extracellular matrix, (C) cytoplasmic inclusions (D) chloroplast. NOTE: Not all cellular organelles can be seen in this image.

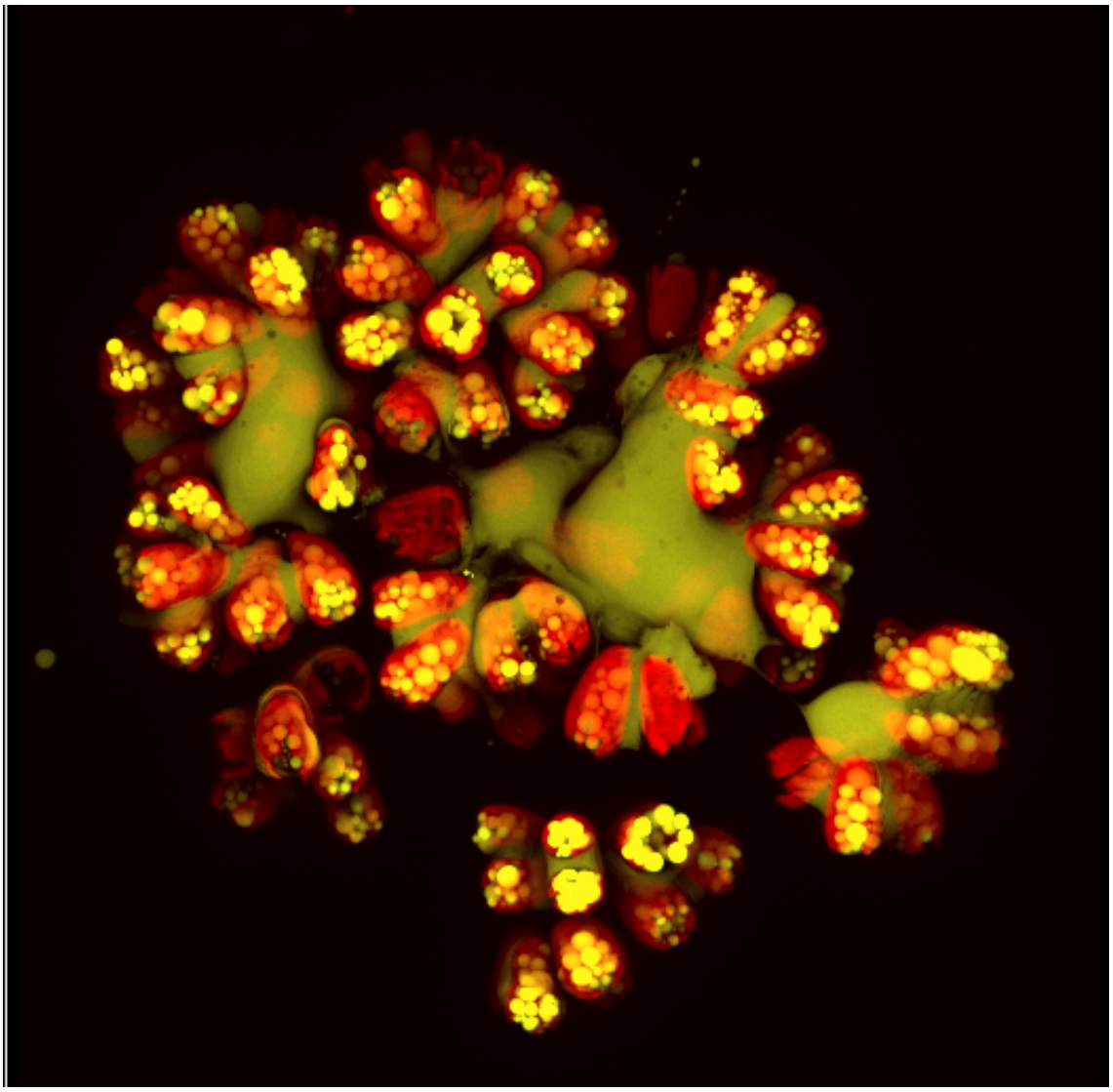


Figure 4 Confocal image of *Botryococcus braunii* "Guadeloupe"
B. braunii colony stained with boron-dipyyromethene. The excitation wavelength was 488 nm. Intracellular hydrocarbon containing-vesicles and the hydrocarbons in the extracellular matrix are visualized in yellow/ green. Chlorophyll autofluorescence (the chloroplasts) are visualized in red. Images courtesy of Dr George Littlejohn (University of Exeter).

1.4.3 Hydrocarbon production

Race A yields 5- 20% algal dry weight of straight chain odd-carbon-numbered n-alkadienes and trienes (Metzger *et al*., 1985a). L- race strains produce lycopadeine, a tetrapadeine (Metzger *et al*., 1990) at a low cellular content of <3% (Metzger and Casadevall, 1987). B-race strains produce 25- 40% algal dry weight of triterpenoid hydrocarbons of C_nH_{2n-10}, n= 30- 37, which are termed botryococcenes (Ishimatsu *et al*., 2012; Metzger *et al*., 1985b). The most abundant triterpenes produced in Race B are typically di-methylated and tetra-methylated botryococcenes. However, under variable culture conditions tetra-methylated squalenes and other derivatives of C_{31}-C_{37} botryococcenes are also produced in lesser quantities (Metzger *et al*., 1985b). Lipids are localised in one of two locations within the cell or colony; the aliphatic outer cell wall or the intracellular oil body. Approximately 7 % of botryococcenes are intracellular, with the remainder being secreted into the ECM (Largeau *et al*., 1980; Wolf *et al*., 1985).

In B race *B. braunii* the degree of methylation of hydrocarbons is dependent on spatial segregation. It is known that C_{30} botyrococcene is produced within cells and modified by methylation, which can occur in the intracellular or extracellular domain. It has been suggested that maturation of intracellular Botryococcenes to C_{34} and higher occurs during transport from the intracellular to the extracellular compartment (Achitouv *et al*., 2004; Metzger *et al*., 1987; Weiss *et al*., 2010). The processes by which botryococcenes are transported and methylated are thus far unconfirmed. The B race strains accumulate overall a larger volume of hydrocarbons, most notably C_{34} methylated botryococcenes, that are more readily extracted and converted by hydrocracking and distillation into high octane gasoline, kerosene and diesel fuels than those of the A and L races (Banerjee *et al*., 2002; Eroglu *et al*., 2010; Niehaus *et al*., 2012).

For the purposes of fuel production, the main goal is to divert the largest fraction of photosynthetic carbon towards the molecular pathways for terpenoid production. This is the opposite of typical carbon-partitioning patterns observed in photosynthetic systems, where 80- 85 % of carbon is directed towards production of biomass and only 5% towards terpenoid pathways. It is proposed that ~ 20% or higher product to biomass carbon partitioning ratio is required within a high productivity photosynthetic system for economical biofuel production. However the *B. braunii* “Showa” B race strain partitions carbon in favour of biofuels production with terpenoid and fatty acid synthesis monopolizing carbon flux (45% and 10% respectively). However, this leads to slow biomass accumulation (Eroglu & Melis, 2010; Eroglu *et al*., 2010). If biomass

productivity could be increased, whilst maintaining the carbon partitioning ratio the major limiting factor to the commercial production of biofuels from *B. braunii* would be removed.

B race strains have been isolated from temperate waters and as such can be cultured at reduced energy expenditure in comparison with the tropical A and L races (Largeau *et al*., 1980). For this reason, coupled with its prolific C_{34} botryococcene production and favourable carbon-partitioning ratio, *B. braunii* race B strain, "Guadeloupe" is of particular commercial interest and was chosen for this study.

1.4.4 Molecular pathways of hydrocarbon synthesis

Methylated botryococcenes are produced alongside similar molecules, methylated squalenes by B- race *Botryococcus* strains. Botryococcenes and squalenes are derived from the isoprenoid pathway, precursors for which are formed via either the mevalonate (MVA) or methylerythrol phosphate 1-deoxy-D-xylulose 5-phosphate (MEP/DOXP) pathway. Generally both MEP/DOXP and MVA pathways are present in algae, as in higher plants where each have a specific role dependent on localization within the cell (Lichtenthaler, 1999). The cytosolic MVA pathway synthesises sterols and triterpenes whereas the plastidic MEP/DOXP pathway forms isoprenes. However, Chlorophytes such as *Scenedesmus obliquus*, *Chlamydomonas reinhardtii* and *Chlorella fusca* appear to only have the MEP/DOXP pathway (Disch *et al*., 1998; Sasso *et al*., 2011; Schwender *et al*., 2001; 1996). Previous studies of *B. braunii* also suggest utilization of the MEP/DOXP pathway (Molnár *et al*., 2012).

Both squalenes and botryococcenes have a backbone formed from the condensation of two C_{15} farnesyl diphosphate residues (Metzger 1990). It is known that in *B. braunii* race B, two farnesyl diphosphate molecules are the precursors for pre-squalene diphosphate (PSPP) synthesis, catalysed by squalene synthase- like 1. Squalene synthase- like 2 was characterised as the catalyst for conversion of PSPP into squalene. Squalene synthase- like 3 converts PSPP into C_{30} botryococcenes. However the downstream molecular pathway of maturation of C_{30} botryococcenes into methylated C_{34} and higher molecules remains uncharacterized (Niehaus *et al*., 2011).

The spatial, molecular and temporal organization of botryococcene biosynthesis in *B. braunii* are not fully characterized.

1.5 Circadian clocks

1.5.1 Importance of circadian control

Day / night cycles are a fundamental feature of the environment and throughout evolution, organisms have adapted to these repeating, predictable rhythms. Daily cycles in the behaviour and physiology of an organism are termed circadian rhythms and a network of transcriptional- translational feedback loop (TTFL) mechanisms radiating from a core clock controls them. Original TTFL models were fairly elementary (Alabadi, 2001) however they have become increasingly more elaborate as research advances (Hsu *et al*., 2013; Locke *et al*., 2006; 2005; Pokhilko *et al*., 2012; Pruneda-Paz *et al*., 2009; Rugnone *et al*., 2013; Zeilinger *et al*., 2006). The oscillatory networks have a cell- autonomous nature in both unicellular (Lakin-Thomas and Brody, 2004) and multicellular organisms (Giebultowicz *et al*., 2000; Plautz, 1997; Welsh *et al*., 1995). Furthermore circadian rhythms are sustained in the absence of environmental cues, for example when an organism is placed under constant environmental conditions, demonstrating that they are endogenous. Whilst the molecular components may vary between species, the mode of action is conserved; a core set of clock genes code for proteins that regulate the expression of clock output genes and pathways throughout the genome (Lowrey & Takahashi, 2004). Signals from the environment are directly or indirectly received by components of the clock, which is entrained to the cycling conditions. The self-contained molecular oscillator of unicellular organisms exerts control over multivariate cellular processes, tissue- specific clock function can be generated in multicellular organisms by partitioning clock circuitry among different cell types (Bell-Pederson *et al*., 2005).

The ubiquitous nature of circadian rhythms and their molecular mechanisms have been defined in plants, fungi, cyanobacteria and vertebrates (Harmer *et al*., 2001). There is evidence to suggest that clocks evolved in nonheterocystous cyanobacteria 3.8 million years ago (Ditty *et al*., 2003) indicating truly pervasive selection pressures on species- wide critical biological processes. Whilst the TTFL mechanism is retained in clocks across Kingdoms, clocks are comprised of different interacting gene groups, although some genes are shared between phyla. Differences in gene sets between Kingdoms indicate distinct evolutionary origins (Young & Kay, 2001). The distinct yet pervasive evolution of circadian control indicates a fundamental and ancient driver for their development.

The escape from UV hypothesis is a logical explanation for conservation of circadian clocks and an obvious example of the need for temporal organisation in most organisms. The vulnerability of DNA to UV and photo-oxidative damage during S-

phase of the cell- cycle demonstrate the necessity to limit this process to hours of darkness only (Pittendrigh, 1993; Rosato & Kyriacou, 2002). Regardless of the particular environmental pressures conferring conservation of clocks, synchronization of processes to an “internal” biological day and likewise for night with the external world enables organisms to anticipate & capitalize on daily environmental changes (Lowrey & Takahashi, 2004; Pittendrigh, 1993).

1.5.2 Circadian control of metabolism

The control exerted over processes within organisms by the circadian clock is extensive. In higher plants transcript profiling has revealed a sophisticated network circadian scheduling of metabolic processes. An example of which is the timing of peak expression of 23 enzymes involved in the phenylpropanoid pathway, just before dawn (Harmer, 2000). Products of the phenylpropenoid pathway include antimicrobial compounds, UV protectants, flower pigments and cell wall components (Camera *et al*., 2004). Cold and stress- induced genes show peak expression at the end of the day, as do several involved in lipid modification. Genes involved in sugar metabolism and transport also demonstrate circadian patterns of rhythmicity (Harmer, 2000). In the Grey Poplar, *Populus* X *canescens,* isoprene biosynthesis is under circadian influence (Loivamaki *et al*., 2006).

In domesticated crops circadian control is involved in processes important to agriculture. We know that, in higher plants, different metabolic products are produced at different times of the day and factors contributing to product quality such as flavour and aroma are metabolized under control of the circadian clock. An example of this is the temporal regulation of *PhCCD1* transcripts, critical to a pathway involved in fragrance in Petunias (Simkin, 2004). Microarray analysis of the grape, *Vitis vinifera*, identified circadian variations in photocycle and thermocycle and showed that berry tissue features could rely on specific circadian molecular oscillations (Carbonell-Bejerano *et al*., 2014; Deluc *et al*., 2007). Therefore, harvesting at the most appropriate time increases crop value.

Chlamydomonas reinhardtii, a Chlorophyte, exhibits clock- controlled rhythms in chemotaxis, phototaxis, cell division and UV sensitivity (Mittag, 2005; Mittag & Wagner, 2005), nitrogen uptake and nitrate reductase activity (Pajuelo *et al*., 1995) and DNA supercoiling (Salvador *et al*., 1998). Furthermore targets for a circadian controlled translational regulator protein in *C. reinhardtii* include genes involved in carbon and nitrogen metabolic pathways (Zhao *et al*., 2004).

In the cyanobacterium *Synechococcus* cell division and transcription of virtually the entire genome, is under circadian control (Liu *et al*., 1995; Wijnen & Young, 2006). As the transcriptome of an organism dictates the active molecular pathways, global transcriptome control by the circadian clock translates to control of the metabolome also.

Although, the algal clock and daily oscillations in metabolism are not thoroughly characterized in algae, it can be hypothesized that as the ancestor of organisms in which such connections are well documented, that they are present.

1.5.3 Circadian clock architecture in photosynthetic organisms

In higher plants such as *Arabidopsis thaliana*, a family of transcription factors, including the closely related Late Elongated Hypocotyl (LHY) and Circadian Clock Associated-1 (CCA1) (Schaffer *et al*., 2001; Wang & Tobin, 1998) share a sequence motif within their DNA binding domain and specificity for the evening element (EE) DNA sequence (Rawat *et al*., 2009). Simultaneous peak transcript expression occurs for *LHY* and *CCA1* just after dawn and protein levels follow suit lagging by approximately 2 hours. Transcript levels of four pseudo-response regulators (*PRR*) subsequently peak sequentially after *LHY* and *CCA1* transcripts, with *PRR9* being the first early in the morning and *PRR5*, *PRR7* and *PRR1*. PRR1 is more commonly referred to as Timing of CAB-1 (TOC1) (Matsushika *et al*., 2000; Strayer *et al*., 2000). PRR proteins are expressed with a phase lag with respect to their transcripts (Fujiwara *et al*., 2008). Towards the end of the day peak TOC1 expression coincides with that of LUX, Early Flowering (ELF) 3 and 4. Assembly of the Evening Complex (EC) occurs from the DNA-binding component, LUX, ELF3 and ELF4. ELF3 and ELF4 are recruited to target promoters (Chow *et al*., 2012; Dai *et al*., 2011).

TOC1 protein, alongside the PRR proteins, represses transcription of cca1 and lhy proteins by binding of the TOC1 conserved CCT domain to the promoters of CCA1 and LHY (Gendron *et al*., 2012; Huang *et al*., 2012) (Figure 5). There is uncertainty over the mechanism for CCA1 and LHY activation. The presence of LUX binding sites in the TOC1 promoter (Dixon *et al*., 2011; Helfer *et al*., 2011) and up-regulation of TOC1 in EC mutants (Dixon *et al*., 2011; Fowler *et al*., 1999; Helfer *et al*., 2011; Kikis *et al*., 2005; Kolmos *et al*., 2009) has led to the hypothesis that the EC down- regulates the expression of TOC1, thus lifting repression on CCA1 and LHY expression.

Only one well-characterized circadian clock model exists for the green alga (Figure 6). *Ostreococcus tauri* is a marine picoeukaryotic Prasinophyte. The *O. tauri* circadian clock is a much-reduced version of that of the higher plant, comprising a negative feedback loop involving a TOC1-like and a CCA1/LHY-like component.

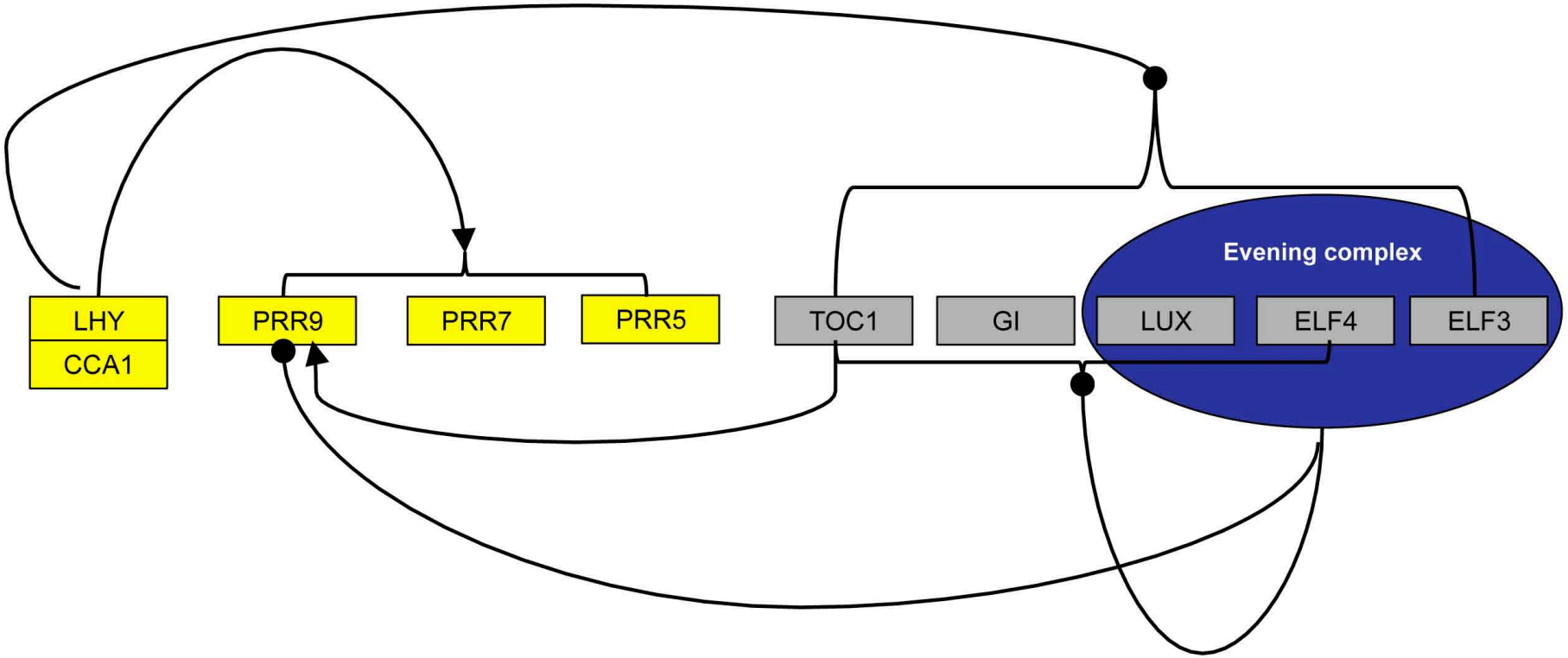


Figure 5 *Arabidopsis thaliana* circadian clock model
A model of the higher plant clock based on that proposed by Pokhilko *et al*, 2012. Proteins only are shown for simplicity. Morning phased elements are shown in yellow boxes and evening phased elements in grey boxes. The association of LUX, ELF4 and ELF3 into an 'evening complex' is denoted by a blue oval. Transcriptional regulation is shown by solid black lines, terminating in a solid circle representing down- regulation or a solid arrowhead for up regulation.

Throughout the day, *O. tauri TOC1* transcription is suppressed by CCA1 via binding to an evening element in the *TOC1* promoter region. *CCA1* transcription is activated by TOC1, which is in contrast to the recently revised *A. thaliana* interaction between TOC1 and *CCA1*. The specifics of the activation of *CCA1* by TOC1 have note been defined but are supported by over-expression/ anti-sense experimental data (Corellou *et al*., 2009). Expression of *O. tauri* TOC1 peaks at dusk, similar to that of *A. thaliana* TOC1. CCA1 transcript abundance begins to accumulate shortly after TOC1 increases and persists until late night, decreasing a few hours before dawn. Only the central clock is described in *O. tauri* and additional feedback loops required for entrainment to different photoperiods via gated response light input pathways remain un-elucidated (Corellou *et al*., 2009).

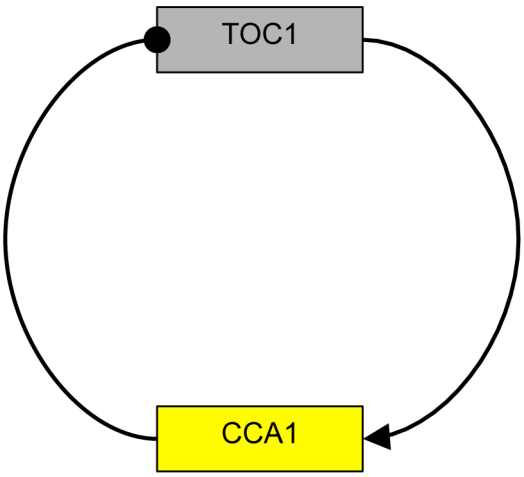


Figure 6 *Ostreococcus tauri* circadian clock model
The proteins of the simple clock mechanism of *O. tauri* are shown with the morning component CCA1 in yellow and the evening component TOC1 in grey. Transcriptional regulation is shown by solid black lines terminating with a solid circle for down- regulation and a solid arrowhead for up- regulation.

The genes, *kaiA*, *kaiB* and *kaiC* are central to the cyanobacterial circadian clock (Ishiura *et al*., 1998). Under continuous light *KaiBC* mRNA and KaiC protein levels oscillate in a circadian fashion (Xu *et al*., 2000). Patterns of rhythmic phosphorylation of these clock components are observed and are proposed to play a role in the maintenance of circadian rhythmicity. The rhythmic phosphorylation of KaiC is promoted by KaiA and repressed by KaiB, and maintained even when transcriptional regulation is eliminated (Kondo, 2007; Tomita *et al*., 2005). Unexpectedly, circadian oscillations in gene expression are observed even when KaiC is artificially maintained in a highly phosphorylated state (Kitayama *et al*., 2008). Thus it is proposed that the regulation of *KaiBC* transcription plays an important role in the cyanobacterial clock as well as post-translational modification via phosphorylation (Zwicker *et al*., 2010). However a definitive model for the cyanobacterial circadian clock is yet to be proposed.

1.6 Identification of clock components

The molecular components of the circadian clock in different organisms have largely been identified by forward genetic screening, *i.e.* the isolation of mutants with impaired circadian function leading to identification and functional characterization of the genetic basis of the mutation. This approach, reinforced with systems modelling, enabled, in *A. thaliana*, the description of approximately 30 clock-associated genes (Nagel & Kay, 2012; Nakamichi, 2011; Pokhilko *et al*., 2012). Despite intense research not all components of the *A. thaliana* clock have yet been unambiguously identified. Moreover, key elements of both output and input pathways remain controversial due to genetic redundancy (Farinas & Mas, 2011; Nakamichi, 2005; Salome, 2005) and compensation whereby knock- down of a component leads to up- regulation of paralogs (Baggs *et al*., 2009). Even though the *A. thaliana* clock may not be fully characterized, it is the most documented plant clock to date and sufficiently well described to be used as comparators against which clocks from non-model plant species may be mapped.

The intricacies of intercellular clock circuitry in multicellular organisms require an experimental set up that can encompass and disentangle a vast and labyrinthine data set (Bell-Pederson *et al*., 2005). Systems biology or omic studies – those that gather substantial data sets from a biological system, *e.g.* genes (genomics), transcripts (transcriptomics) and metabolites (metabolomics) – are equipped to unravel the aforementioned complexities. Whole transcriptome study by microarray demonstrated that rhythmic patterns at the whole plant level are induced at a molecular level (Harmer, 2000; Schaffer *et al*., 2001). Recent advances in genome- wide technology

have allowed comprehensive studies to show that the molecular signatures, from gene expression, protein level and activity and metabolite profiles, of rhythmic behaviours and processes change over a 24- hour period (Chow & Kay, 2013). Splice variants detected by massively parallel short read RNA sequencing (RNA-Seq) revealed whole new level of circadian regulation in plants (Filichkin *et al*., 2010; Henriques and Mas, 2013). RNA and DNA sequencing technologies have revealed that target genes for core circadian transcription factors are highly enriched for metabolic pathways and that one third of all expressed *A. thaliana* genes are clock regulated- a much larger estimate than that preceding (Covington *et al*., 2008; Koike *et al*., 2012).

Systems biology is dependent on high throughput technologies, such as Next Generation Sequencing (NGS), otherwise termed “massively parallel sequencing”, which has evolved in a relatively short space of time since the sequencing of the human genome.

1.7 DNA sequencing technologies

1.7.1 Chemical and chain termination sequencing

The first successful DNA sequencing methods, chemical sequencing and chain termination sequencing, were published in the 1970s and had broadly similar protocols. Chemical sequencing utilized 4 reactions that preferentially cleaved terminally labeled DNA molecules at positions of guanines, adenines, cytosines or thymines repetition. The fragmented resolved by size on an electrophoretic gel and the sequence of the labeled DNA sequence could be determined from the pattern of bands (Maxam and Gilbert, 1977). Chain termination sequencing, the alternative method, utilised incorporation of 2',3' dideoxy- and arabino- nucleoside analogues specific to each base as a DNA synthesis terminating step in 4 separate reactions. The reaction products were then resolved on an electrophoretic gel and the pattern of radioactive bands used to determine the nucleotide sequence (Sanger *et al*., 1977). Chain termination sequencing rapidly superseded chemical sequencing.

The present day “Sanger” technology is based on a modified chain termination method. A single strand of DNA is amplified in four separate reactions by addition of three of the four dNTPs (dATP, dTTP, dGTP or dCTP) and a modified fluorescently labeled strand- terminating ddNTP. Following rounds of amplification the resulting fragments are sequenced by capillary gel electrophoresis according to length, detection and recording of dye fluorescence. The data are output as fluorescence peak chromatograms from which the sequence can be read. A further development of this

technology is paired- end sequencing whereby both ends of a DNA fragment are over-sampled giving rise to paired reads, which as a result of the over-sampling overlap allowing reconstruction of contiguous sequence using whole- genome assembly algorithms (Mutz *et al*., 2013).

1.7.2 PCR-based sequencing

The Roche 454 FLX Pyrosequencer was the first of three NGS platforms in widespread use to become commercially available. Library fragments are fused to agarose beads via oligonucleotides specific to the 454- adapter sequences on the fragment library and beads are arrayed in a picotiter plate, one bead per well. During subsequent rounds of amplification the four pure dNTPs are introduced individually and release of a pyrophosphate upon each dNTP addition initiates a downstream reaction that ultimately leads to the emission of light from luciferase. Between each nucleotide incorporation event there is an imaging step, and sequence data can therefore be inferred for each library according to the fixed location defined by well position (Mardis, 2008).

The Applied Biosystems SOLiD™ (Sequencing by Oligonucleotide Ligation and Detection) sequencer uses an approach combining that of the Roche 454 library- to-bead binding system with a unique technique employing DNA ligase. Fragments of DNA are bound to adapters on the bead surface and replicated across the bead surface by PCR within a water “microenvironment” contained within an oil droplet, one bead per droplet (emulsion PCR). Universal primers are bound to the fragments, followed by addition of four di- base probes, with colour- coded fluorophores attached. Each one of the di- base probes may encode one of four known pairs of bases. These di-base probes compete to ligate to the universal primer (n) and specificity is achieved by interrogation of every first and second base in each ligation reaction. Colour- coded fluorescence is detected and imaged before un-extended fragments are capped and cleavage of the fluorophore occurs to prepare for the next addition of di- base probes. A user-defined number of rounds of ligation, excitation, imaging and cleavage are performed before primer reset is performed. During primer reset the above outlined cycle is repeated 5 times with primers of the length n- 1, n- 2, n- 3 and n- 4, creating a base shift of 1 in the available fragment to bind to for the di- base probes for each reset. Image analysis & sequence delineation follow, entailing assignment of possible di-nucleotides at each read position and double interrogation based on the subsequent di-base probe binding at that position in the following rounds of primer reset (Mardis, 2008).

1.7.3 Illumina sequencing

The Illumina Genome Analyzer followed the Roche 454 FLX platform. Illumina technology uses the sequencing-by-synthesis approach (Figure 7). Libraries of randomly fragmented sample DNA or cDNA are bound to the surface of an eight channel, sealed glass flow cell via adapter sequences that are complementary to oligomers fixed to the flow-cell glass during manufacture. Depending on the sample, one or several libraries may be loaded into a single channel. In the latter case, termed "multiplexing", libraries are barcoded with different, known leader sequences during preparation and therefore can be run in the same lane to minimize cost. To ensure a detectable signal, the individual DNA fragments that compose the library are first immobilised in the flowcell and multiplied by solid phase bridge amplification (a form of PCR) to generate clusters of cloned double-stranded fragments. The DNA is then denatured and sequenced, base-by-base using modified nucleotides each with a base specific fluorescent signal. After each round of synthesis, the reaction is chemically blocked by the presence of a 3'-OH group, and the incorporated nucleotides imaged. The clustered configuration that represents a cloned library fragment results in resolved spots which colours represent the incorporated bases. Subsequently the 3'-OH group block is removed from the incorporated base and the synthesis-imaging-regeneration steps repeated (Mardis, 2008).

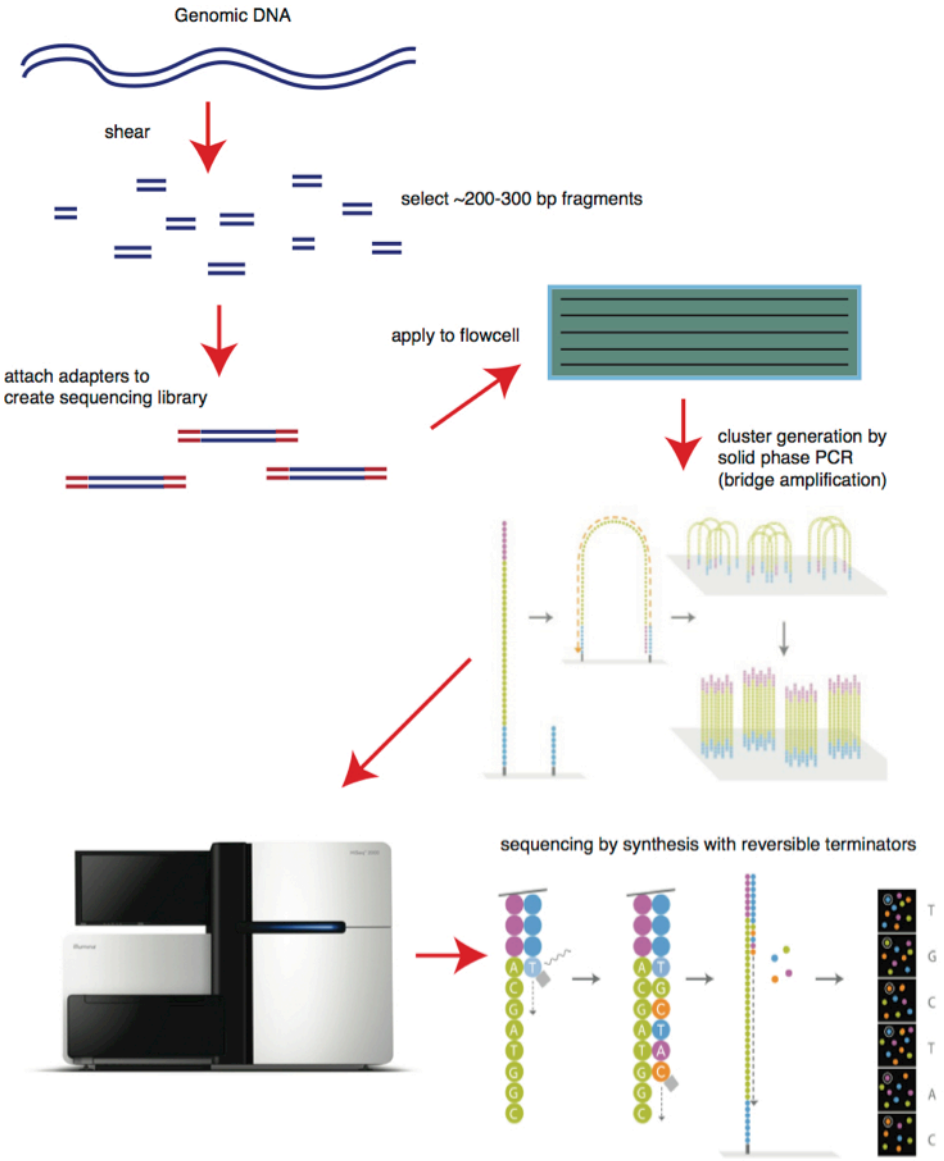


Figure 7 Diagramatic representation of Illumina sequencing

Overview of library preparation and sequencing technology involved in Illumina sequencing. Output is shown as a digital colour-dot read-out, which is translated into notated sequence by image analysis software.

The image is courtesy of (http://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/).

1.8 RNA-Seq

NGS technology has revolutionized transcriptomics as well as genomics, with recent advancements enabling massively parallel sequencing of mRNA-derived cDNA, termed RNA-Seq (Wang *et al*., 2008; Yassour *et al*., 2009). The term RNA-Seq is not literal; library preparation involves a reverse transcription step, allowing the NGS platforms described above to be directly applied to the sequencing of mRNA via cDNA. RNA-Seq is essentially a massively parallel and internally cross-referenced quantitative PCR and is a more accurate measure of absolute expression and can detect a broader range of expression levels than microarray technology. Further, splice variants, novel isoforms of genes and closely related paralogous genes can be detected and distinguished between by RNA-Seq. No prior knowledge of genomic sequence is required for the application of RNA-Seq, as in the case of microarray techniques so the technology lends itself well to the study of non-model organisms (Fu *et al*., 2009; Marioni *et al*., 2008; Wilhelm *et al*., 2008).

Applications of RNA-Seq include relative expression and absolute quantification of mRNA, identification of novel transcripts, isoforms and protein-binding sites, and characterization of non-coding RNA (Mutz *et al*., 2013). Differential expression analysis can also be performed on the raw count data generated by RNA-Seq, to infer whether or not abundance of transcripts is statistically different between samples.

1.8.1 The application of RNA-Seq to circadian biology and metabolism studies

In higher plants, the circadian clock influences the expression of 40-60% of the entire genome (Harmer; (Doherty & Kay, 2010). This control generates networks of co-expressed and anti-expressed genes and enzymatic or metabolic pathways. RNA-Seq analysis of samples taken over a sufficiently resolved time-course allows investigation of the differences in temporal expression of genes, which can be compared across different entrainment conditions to elucidate, albeit indirectly, circadian and photoperiodic controls on metabolic pathways. Gene expression data generated by RNA-Seq can also be used to qualitatively group genes that show similar patterns of expression (in this case, the pattern would be temporal), thereby reproducing networks of co- and anti-expressed pathways (Iancu *et al*., 2012; Tang *et al*., 2011).

RNA-Seq data can also be used to identify novel core clock components directly, by in depth, comparative analysis of transcripts that display oscillatory expression profiles in constant conditions (Boothroyd *et al*., 2007). Additionally highly intricate

clock gene regulation by alternative splicing can be detected by RNA-Seq (Chow *et al*., 2012).

Thus, the temporal metabolic activity of organisms can be thoroughly investigated by comprehensive analysis of whole transcriptome sequence data (Birol *et al*., 2009; Garber *et al*., 2011).

1.8.2 De novo *transcriptome assembly of* B. braunii

Whole transcriptome analysis is an intricate and time-consuming process involving diverse and complex analyses. One Illumina HiSeq sequencing lane can produce up to 200 million, 100 bp reads; this equates to approximately 50 GB of data. Analysis of these data required improved algorithms capable of managing such vast tasks, compared to more conventional genomics based on Sanger sequencing. Raw sequence reads must be assembled into transcripts. These transcripts are either mapped to an existing reference genome or transcriptome or used to generate, *de novo*, a previously uncharacterized transcriptome. Reads are then quantified and differential expression inferred by statistical analysis. Further advancements in technology have, in recent years, reduced some of the challenges posed by whole transcriptome sequencing due to the open-source availability of sophisticated and well-documented bioinformatic and statistical tools (Chu & Corey, 2012).

1.9 Project Plan

Microalgae are a promising source of advanced or “third generation” biofuel, either by direct biomass, lipid or hydrocarbon production, or indirectly as a source of biosynthetic pathways for industrially relevant compounds. For example, the terpenoid farnesene is currently produced industrially in Brazil using synthetically engineered, heterotrophic microbes as an alternative biochemical and fuel precursor, by the USA corporation Amrys.

Botryococcus braunii is particularly interesting because it synthesizes and secretes up to 80% of its dry mass as long-chain (C_{30}-C_{34}) hydrocarbons that are derived from the terpenoid pathway and collectively termed botryococcenes (Banerjee *et al*., 2002). Although the biochemistry of botryococcene synthesis has been largely elucidated (Sato *et al*., 2003), few genes encoding the enzymes involved in botryococcene production have been characterized (Ioki *et al*., 2012) and no information regarding the entire genetic basis and molecular coordination of the

botryococcene pathway is available.

In the literature, botryococcene biosynthesis is typically regarded as dependent on growth. However, that assertion may be overly crude because in microalgae, as in higher plants, the timing of key metabolic processes such as sugar allocation or lipid production may be controlled both by photoperiod (day-length) and through cellular rhythms generated by the endogenous circadian clock (Johnson, 2001; Kiyota *et al*., 2006; Mittag, 2001). Precise knowledge of the timing of hydrocarbon production in *B. braunii* may therefore be important to determine whether there exists a daily window during which harvesting *B. braunii* for oils is most productive.

This project aims to fill these gaps in our knowledge first by elucidating the molecular pathway for botryococcene biosynthesis and second, by investigating whether the circadian clock or photoperiod have an influence on the expression of the genes controlling hydrocarbon biosynthesis.

A functional transcriptomic approach was adopted to achieve these aims. *B. braunii* were entrained in 12 h light and 12 h dark photoperiods, and either maintained in that same photoperiod or transferred to constant light (*i.e.* circadian free-run). cDNA libraries were generated at different points in the diel cycle and sequenced using Illumina NGS, paired-end RNA-Seq (Figure 8). Transcriptomes from each time-point were amalgamated to generate the entire diel transcriptome, which was then assembled *de novo*, annotated and analyzed using a variety of bioinformatics tools (Figure 9). The dataset was validated by resolving and modelling the core elements of the *B. braunii* circadian clock, and comparing it to previously described clock networks in *A. thaliana* and *Ostreococcus tauri*. Following validation, the genes encoding the botryococcene pathway were identified, and show some level of genetic redundancy. To ascertain the abundance of all transcripts at all time points, and thereby the influence of either photoperiod or circadian clock on gene expression, the time-point specific reads were aligned to the amalgamated transcriptome. These data confirmed the role of the MEP/DOXP pathway for botryococcene synthesis and that this pathway is largely under circadian and photoperiodic control.

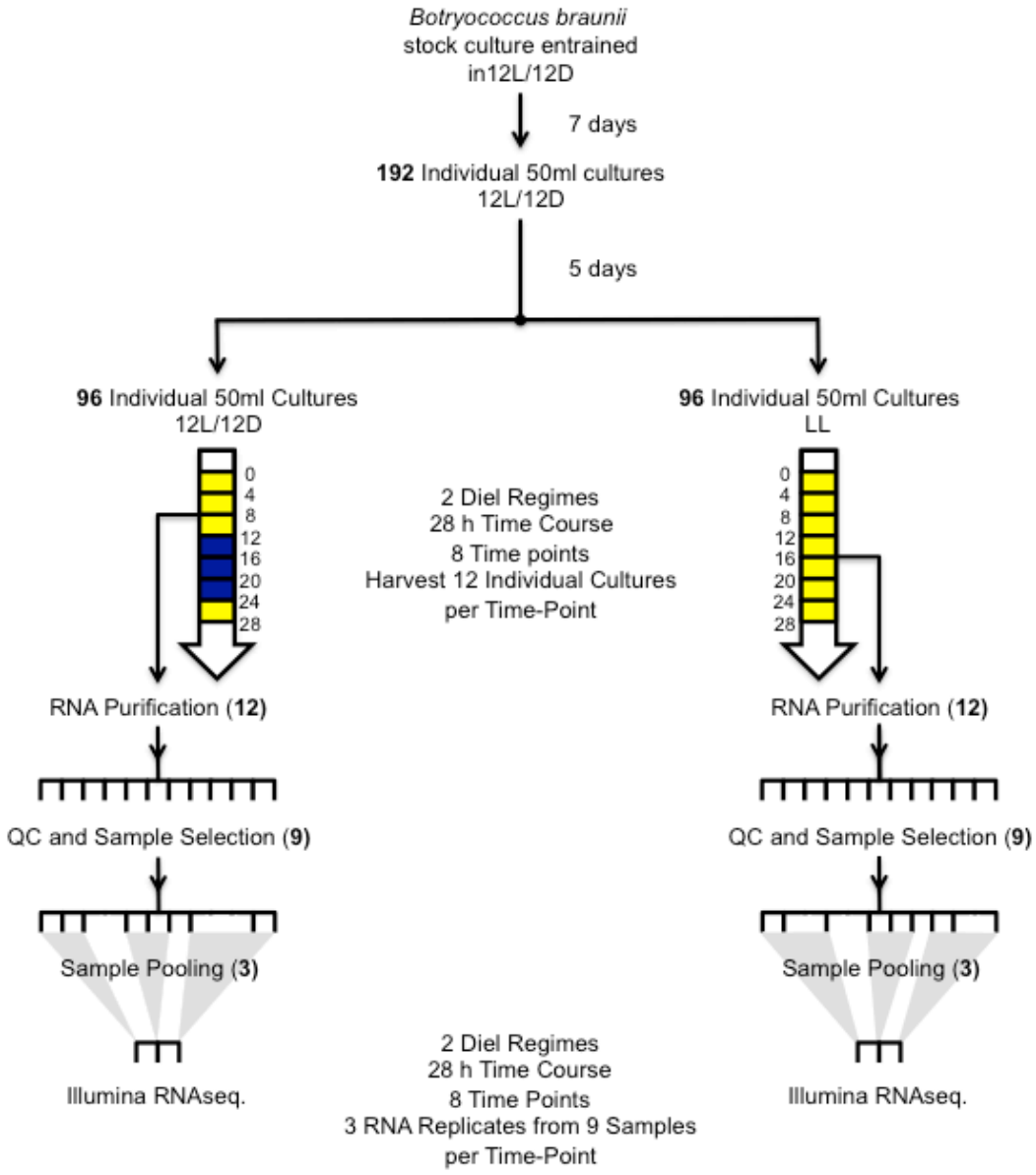


Figure 8 Schematic representation of sample processing
Diagram showing the processes from *B. braunii* sample harvesting to prepared Illumina cDNA libraries.

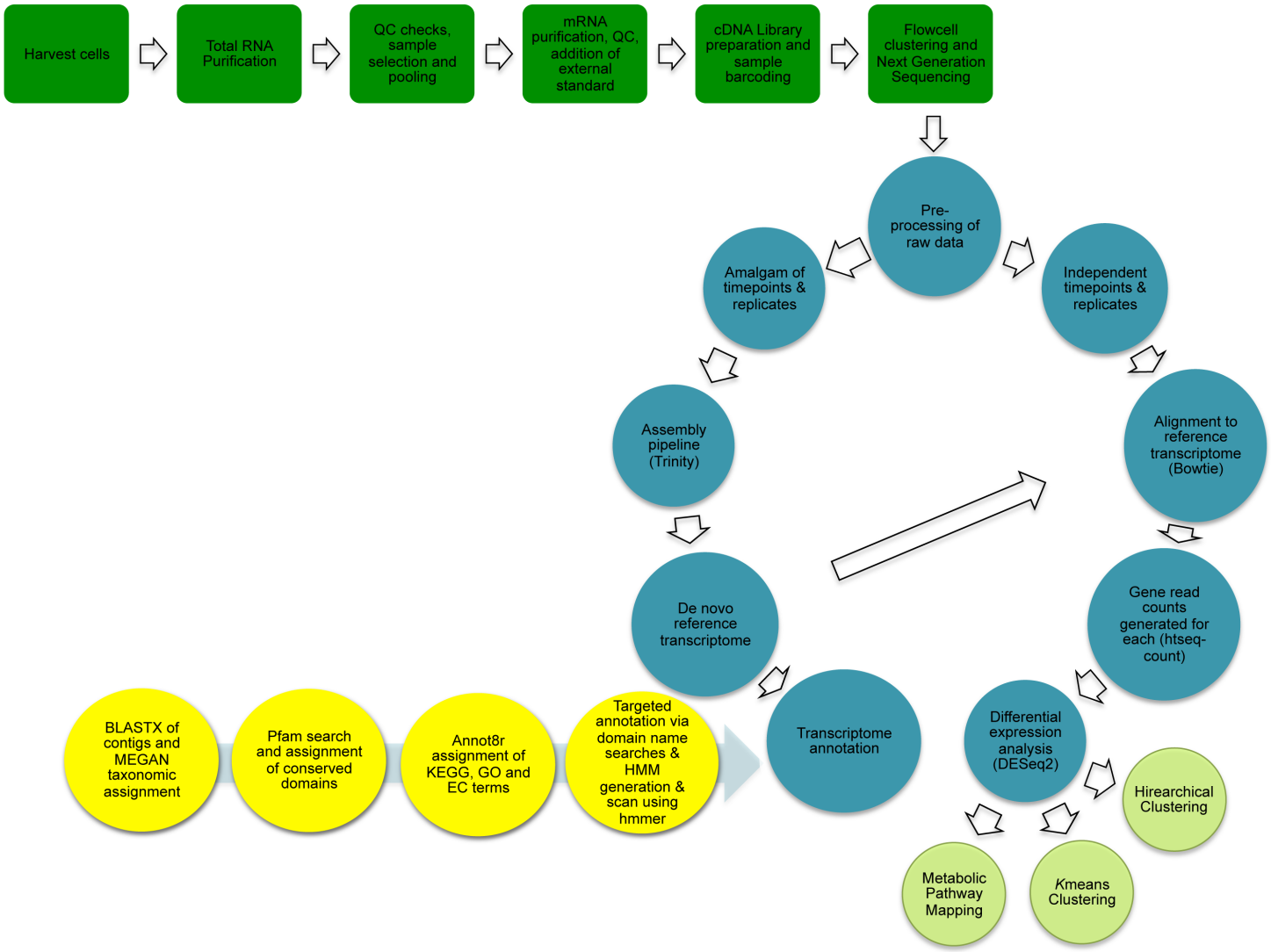


Figure 9 Workflow showing experimental and bioinformatic procedures

Practical laboratory methods (green) through to bioinformatic analysis (teal blue) used to investigate the transcriptomics of *B. braunii* over a 28-hour time series under two different photo-regimes. Chapter 3 describes the generation and annotation of the *Botryococcus braunii* transcriptome (green, teal blue and yellow) sequence. Faded aspects are addressed in Chapters 4 and 5.

CHAPTER 2

GENERAL MATERIALS AND METHODS

Unless otherwise stated in this chapter all water used throughout the thesis was sterile ultrapure 18 MOhm MilliQ water (Millipore). Consumables and reagents for general laboratory practice were obtained from Sigma-Aldrich or Fisher-Scientific and were of analytical grade as a minimum standard.

Glassware was prepared for analytical applications by removal of organic residues in a Miele Professional G 7883 CD laboratory glassware washer using programme F. Glassware for algal culture and RNA applications was autoclaved at 121 °C for 20 min at 15 psi. All glassware was thoroughly dried in a drying oven at 45 °C.

2.1 Media preparation

2.1.1 Modified Chu 13

The medium used was modified Chu 13, supplemented with sodium selenite (400 mg l^{-1} KNO_3, 200 mg l^{-1} $MgSO_4 \bullet 7H_2O$, 108 mg l^{-1} $CaCl_2 \bullet 2H_2O$, 104.8 mg l^{-1} K_2HPO_4, 20 mg l^{-1} FeNaEDTA , 2.86 mg l^{-1} H_3BO_3, 1.9 mg l^{-1} Na_2O_4Se, 1.3 mg l^{-1} $MnSO_4 \cdot 4H_2O$, 220μg l^{-1} $ZnSO_4 \cdot 7H_2O$, 90 μg l^{-1} $CoSO_4 \cdot 7H_2O$, 80 μg l^{-1} $CuSO_4 \cdot 5H_2O$, 60 μg l^{-1} $Na_2MoO_4 \cdot 2H_2O$, 10 ul l^{-1} H_2SO_4). Whilst mixing on a stirring plate the pH of the medium was adjusted to 7.5 with KOH before autoclaving for 20 min at 120 °C. Autoclaved sterile $CaCl_2 \cdot 2H_2O$ was added after autoclaving.

2.1.2 Luria Bertani broth agar preparation

Luria Bertani (LB) broth agar was prepared by adding 10 g bactotryptone, 10 g NaCl, 5 g yeast extract and 15 g agar to 1 l MilliQ H_2O. Media was allowed to mix thoroughly on a stirring plate before autoclaving for 20 min at 120 °C.

2.1.3 Yeast Mould agar preparation

Yeast Mould (YM) agar contained, per litre, 20 g agar, 10 g glucose, 3 g yeast extract, 3 g malt extract and 5 g peptone. Media was mixed thoroughly on a stirring plate before autoclaving for 20 min at 120 °C.

2.1.4 mRNA buffers

mRNA binding buffer was made up from 20 mM Tris-HCl (pH 7.5), 1.0 M LiCl and 2 mM EDTA. Bead washing buffer comprised of 10 mM Tris-HCl (pH 7.5), 0.15 M LiCl and 1 mM EDTA. Beads were reconditioned with 0.1 M NaOH solution and stored in 250 mM Tris-HCL (pH 7.5), 20 mM EDTA, 0.1% Tween20, and 0.02 % NaN_3.

2.2 Algal culture conditions

The *Botryococcus braunii* race B Guadeloupe strain used for all experiments was derived from culture obtained from Pierre Metzger at Laboratoire de Chimie Bioorganique et Organique Physique, Ecole Nationale Supérieure. De Chimie De Paris (Metzger, Berkaloff, Casadevall, & Coute, 1985).

Algae were grown in Infors HT Multitron incubators with 5% atmospheric CO_2, at 23°C, and with shaking at 105 rpm. Illumination was provided for a period of 18 h in a 24 h diel cycle unless otherwise specified. Illumination was by Sylvania Grolux 15W T8 lamps. Total light flux density was 70 μmol m^{-1} s^{-1}, measured using a Hansatech Quantitherm Light Meter.

Algae were maintained in pre-stationary phase by weekly subculture. Cells were centrifuged at 12,000 g for 15 min and transferred to a 20 μm sieve. Algae were washed using 5 l of sterile MilliQ water. The algae that remained in the sieve were re-suspended in Modified Chu medium before re-inoculation into flasks such that OD_{680} of the new cultures was approximately 0.2. All of these steps were performed in aseptic conditions in a Scanlaf Mars sterile laminar flow cabinet.

2.3 Monitoring potential contamination of algal cultures

A rapid check for the presence of contaminating organisms in *B. braunii* cultures was performed by observation under a light microscope.

As a more thorough contamination check, once monthly and in preparation for an

experiment, the number of bacterial Colony Forming Units (CFU's) per ml of algal culture was determined. A dilution series from the algal cultures that ranged from 10^{-1} to 10^{-7} was prepared and 100 μl of each dilution spread on to LB agar plates. After incubation at 25°C for 48 h bacterial colonies were counted. If fungal contamination was observed to be present in the preliminary microscopy step, 100 μl of each dilution was also spread on to YM agar and incubated as described above for the bacterial CFU check.

2.4 Absorbance of algal culture

The absorbance of 250 μl of algal culture was quantified at OD_{680} nm (Abs_{680}) in Genetix Petriwell 96- well plates using a Tecan Infinite M200 microplate reader. Modified Chu medium was used as a blank and the values were subtracted from the total culture readings.

2.5 Algal cell harvest

Liquid cultures were filtered through 45 mm GF/C 1.2 μm Whatman filter papers (unless otherwise specified) and the retained algal cells immediately scraped off with a sterile mounted razorblade. For RNA purification or long- term storage, harvested cells were immersed into liquid N_2. Flash-frozen cell pellets were transferred into chilled Greiner CRYO.S™ 2 ml tubes and stored at -80°C.

2.6 Quantification of dry biomass

5 to 10 ml algal cells were harvested on to dried and pre-weighed filter papers as described previously, and dried at 120°C for 24 h before being weighed, returned to the drying oven and re-weighed after a further 24 h. Dry algal biomass was determined as the final total weight of the filter paper plus the algal cells, minus the filter weight and multiplied by the appropriate factor to provide a dry biomass in mg l^{-1}.

2.7 Quantification of chlorophyll

B. braunii cells were harvested on a filter paper as described in section 2.5 Algal cell harvest. The following steps were then performed in the dark. Filter papers containing the freshly harvested cells were immersed in 10 ml methanol and sonicated at 45 kHz in a VWR Ultrasonic Cleaner at room temperature for 20 min or until all traces of green colouration had disappeared from the filter paper. The organic solution was retrieved and centrifuged at 2,000 g for 5 min to pellet cell and filter paper debris. Chlorophyll absorbance of 250 μl of chlorophyll extract was then measured at 650 nm and 665 nm using a Tecan Infinite M200 microplate reader. 250 μl methanol was used as a blank and the values subtracted from the total chlorophyll extract readings. Chlorophyll concentrations were calculated according to the following formulae (Hipkins & Baker, 1986):

Chlorophyll A = $(16.5Abs_{665}) – (8.3Abs_{650})$

Chlorophyll B = $(33.8Abs_{650}) – (12.5Abs_{665})$

Total chlorophyll = $(25.8Abs_{650}) + (4Abs_{665})$

To give chlorophyll in mg l^{-1} the appropriate dilution factor was applied.

2.8 Total RNA extraction from *Botryococcus braunii* cells

Frozen *B. braunii* cells from 50 ml culture at approximately 0.5 OD_{680}, were dispensed into 1 ml Qiazol Lysis Reagent (Qiagen) in Lysis Reagent A tubes (MP Biomedicals) and homogenised at 6.0 m/s for 40 s in a FastPrep-24 bead beater (MP Biomedicals). Cell debris was pelleted by centrifugation at 16,100 RCF at 4°C for 5 minutes. The clarified lysate from below the floating cell debris was transferred to a Maxtract High Density tube (Qiagen) and hand shaken for 15 s with 100 μl 1-Bromo-3-chloropropane (BCP). The lysate was incubated for 3 minutes at room temperature and centrifuged for 15 minutes at 12,000 RCF (4°C) The supernatant was transferred to a fresh tube and vortexed with an equal volume of isopropanol to precipitate nucleic acids. The isopropanol with contained RNA was applied to an RNeasy spin column; from this point onwards the RNeasy Plant Mini Kit (Qiagen) protocol was followed. Briefly the sample was collected in the RNeasy spin column by centrifugation and the flow through discarded. On-column DNase digestion was performed by incubation at room temperature for 15 min with DNase1 solution. The spin column membrane was then washed and the flow through

discarded. RNA was eluted into a clean 1.5 ml reaction tube with 50 µl RNase-free water. RNA was quantified prior to storage (2.11 Quantification and analysis of nucleic acids).

2.9 Addition of external RNA controls

Prior to mRNA purification for cDNA library preparation, 1 μl of a 1/10 dilution of External RNA Control Consortion (ERCC) ExFold RNA Spike- In Control Mix (Ambion Life Technologies) was added and thoroughly mixed with a volume of each sample containing 5 μg RNA before proceeding with mRNA purification.

2.10 mRNA purification

The Dynabeads® mRNA Purification Kit (Invitrogen, UK) was used to purify mRNA from 5 μg total RNA in 100 μl. Briefly; the RNA was heat denatured at 65 °C for 2 min before polyA- tail binding to washed and calibrated oligo$(dT)_{25}$ (Dynabeads) by mixing on a rotating roller for 5 min at room temperature in 100 μl binding buffer. A magnetic tube rack was used to remove the mRNA from suspension, drawing the beads out of suspension until the solution was clear. The mRNA-bead complexes were washed twice with washing buffer B. mRNA was eluted into 10 μl 10 mM Tris-HCl (pH 7.5) by denaturation of the A-T binding at 65 °C for 2 min and placing immediately on to the magnetic rack to separate beads. Samples were placed into a clean chilled RNase- free tube for storage at -80°C.

2.11 Quantification and analysis of nucleic acids

2.11.1 Nanodrop analysis

Nucleic acid (NA) concentration was determined on a THERMO Scientific NanoDrop ND1000 spectrophotometer. The spectrophotometer was initialised using 2 μl nuclease-free water as a blank after which 2 μl of undiluted DNA or RNA samples were applied to the electrode and the concentration quantified in ng/ μl using the NanoDrop 1000 software. Concentration calculation was based upon a modified Beer- Lambert equation; c = A/(E x b), where c is the nucleic acid concentration in ng/ μl, A is the absorbance in AU, E is the wavelength dependent extinction coefficient in ng- cm/μl, b is the path length in cm. The extinction coefficient used for RNA quantification was 40. An automatic path length change of 1 mm to 0.2 mm was incorporated in the measurement cycle, depending on sample concentration.

2.11.2 Agilent Bioanalyzer analysis

Capillary electrophoresis on an Agilent 2100 Electrophoresis Bioanalyzer instrument was used to size separate and quantify RNA or DNA molecules loaded into a microfluidic chip based on their constant mass to charge ratio.

2.11.2.1 RNA analysis

The Agilent RNA 6000 Nano kit was used to process RNA and a microfluidic chip for RNA separation, sizing and quantification by capillary electrophoresis. RNA 6000 Nano Marker, RNA 6000 Nano Dye Concentrate and RNA 6000 Nano Gel Matrix were incubated at room temperature for 30 min prior to use.

The RNA Nano microfluidic chip was loaded with 9 μl Nano Gel Matrix mixed with 1 μl Nano Dye, which was forced into the microchannels by pressure applied to the gel well via a 1 ml syringe. 5 μl of Nano marker was added to all sample wells and the ladder well, followed by 1 μl of samples and 1 μl ladder respectively. The microfluidic chip was vortexed at 2400 rpm and immediately inserted into the Bioanalyzer. Electrodes were immersed into each loaded well to generate a charge across the microfluidic chip, allowing size separation as the RNA and reference ladder migrated through a sieving polymer in glass microchannels interconnecting the wells. Dye molecules intercalated into the RNA molecules allowed for laser- induced detection, the data generated from which was incorporated into gel- like images and electropherograms. Agilent 2100 Expert Software version 1.2 was used to run the PlantTotal RNA Nano assay on an Agilent 2100 Bioanalyzer. RNA concentration was determined based on a calculation using the reference ladder concentration to peak area relationship applied to sample peak area. The microfluidic chip reader was cleaned with RNaseZAP™ (Life Technologies) and rinsed with nuclease free water before and after use.

2.11.2.2 DNA analysis

The Agilent DNA 7500 kit was used to process cDNA samples for analysis on the Bioanalyzer. The prepared DNA Gel Matrix with added DNA Dye Concentrate incubated at room temperature for 30 minutes prior to use, whilst the DNA samples were thawed on ice. According to the manufacturer's instructions and similar to the described RNA protocol

(section 2.11.2.1), the chip was loaded with the Nano Gel Matrix, followed by samples, internal standards and ladder and Agilent 2100 Expert Software was used to run the DNA 7500 assay on an Agilent 2100 Bioanalyzer. The chip reader was cleaned with RNAse ZAP™ (Life Technologies) and rinsed with nuclease free water before and after use.

The Agilent 2100 software created a standard curve of DNA migration time against fragment size from the running of the ladder Concentration of cDNA fragments in the size range of 100- 700 base pairs was determined by manual selection of this range on the output graph and using the software for an automated calculation of concentration in ng/μl. Size and concentration are calculated by aligning (in terms of detection time) the lower and upper well markers with the first and last ladder peaks and a calculation based on derived on the area to concentration relationship with the upper ladder marker.

2.12 Construction of cDNA libraries and sequencing

cDNA libraries were prepared and amplified using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre) and Kapa HiFi Library Amplification Kit (Kapa Biosystems). Briefly, RNA was fragmented before cDNA was synthesised from the antisense strand. cDNA fragments were terminal tagged at both 3' and 5' ends and the library was purified. During the subsequent amplification by polymerase chain reaction, the libraries were barcoded for multiplexing (*i.e.* where multiple samples are loaded per sequencing lane) by replacement of the reverse primer with 1 μl 10 μM Epicentre Index Primer, ensuring that each library within one sequencing lane was identifiable by a different primer. Primers were selected such that at each position of the index code to be read during sequencing, there was at least one A or C and at least one T or G; ensuring accuracy of signal detection and therefore confident delineation between indexed samples. (Primer sequences are given in Table 1 in the Appendix). Finally, cDNA fragment libraries were purified using Agencourt AMpure XP system (Beckman Coulter). Library concentration and size distribution of fragments were determined using an Agilent Bioanalyzer and Expert Software version 1.2, before libraries were diluted to 10 nmol and stored at -20°C.

For sequencing, the libraries were distributed into eight multiplexed groups and each group loaded into one lane of an Illumina 8-lane flow cell at a concentration of 6.5 pmol. Using the TruSeq PE Cluster Kit v3 - cBOT (Illumina), the single cDNA molecules available within each library were covalently bound by complementary adaptors to the flow cell surface and clonal clusters generated using the Illumina cBOT automated system. The flow cell was finally prepared for paired-end sequencing using TruSeq SBS Kit HS v3 (Illumina) and run on an Illumina HiSeq 2000 by the University of Exeter Sequencing Service.

CHAPTER 3

FROM *BOTRYOCOCCUS BRAUNII* CULTURES TO ANNOTATED TRANSCRIPTOME

3.1 Introduction

The slow growth rate of *B. braunii* places severe limitations on its suitability as a feedstock in industrial scale biofuel production. However, opportunities lie in the potential for engineering useful pathways from *B. braunii* into more prolific organisms and the modification of pathways within *B. braunii* to increase production. These routes to increasing the amenability of *B. braunii* as a biofuels feedstock are dependent upon the capture of information pertaining to the sequence of genes encoding for enzymes involved in pathways of interest (Niehaus *et al*., 2011).

Although a *B. braunii* race B genome has recently been sequenced, there is no assembled genome publically available (http://genome.jgi.doe.gov/genome-projects/pages/projects.jsf?searchText=Showa). Delays in the sequencing and successful annotation of a genome can be attributed to the difficulty of obtaining and maintaining axenic cultures of *B. braunii*. Cultures are often contaminated with bacteria (Khatri *et al*., 2014; Lupi *et al*., 1991), which associate with the exopolysaccharide matrix of *B. braunii* colonies (Rivas *et al*., 2010). Bacterial contamination of cultures can result in the generation of the vast majority of genomic sequences originating largely from bacterial sources. The reasons for this are twofold: firstly because of the high number of bacterial cells compared to their algal counterparts, and secondly because the genomic yield per bacterial cell is greater than the yield per algal cell due to an extraction bias, *i.e.* DNA is readily extracted from bacterial cells but less so from plant material because of the tough cell walls. Association of bacterial consortia with the extracellular matrix (ECM) of *B. braunii* colonies renders the removal of contamination from the culture without damaging microalgal cells challenging. Further, it is possible that *B. braunii* may benefit from a symbiotic or commensal relationship with a bacterial consortium for obtaining vitamins and other essential micronutrients, B12 in particular. *De novo* synthesis of B12 is so far undiscovered and long term stability and effects of the axenic culture of *B. braunii* are poorly understood. The slow growth of axenic cultures of *B. braunii* supports the notion that co-occurring bacteria

are essential for long term viability and stability of cultures (Rivas *et al*., 2010). Additionally, there may be benefits to maintenance of bacterial associations such as the exclusion of other harmful contaminating organisms by co-culture with *Rhizobium* spp. and acceleration of the growth rate (Tanabe *et al*., 2012), also provision of CO_2 from respiring bacteria or bio-available nitrogen, phosphate and other inorganic compounds. In contrast, bacterial association may be antagonistic to *B. braunii* health in some aspects through the release of toxic compounds or competition for identical resources (Lupi *et al*., 1991).

Transcriptomic data has the potential to provide insight and understanding on the specific and general metabolic requirements of an organism. For example, as mentioned previously, it could reveal whether or not *B. braunii* needs to acquire B12 directly from the environment, or if they are indeed B12 autotrophs, by the presence or absence of methionine synthase transcripts, the expression of which can be used as an indication that the algae are synthesising their own B12 (Tanabe *et al*., 2013). Such understanding imparts important knowledge of potential effects of long term axenic culture *of B. braunii,* thus feeding back into genomic-based studies.

Inferring the active pathways within an organism and biological process involvement, and molecular function of transcripts within an assembled *de novo* transcriptome, rely on the availability of a dataset of annotated sequences that is as comprehensive as feasible, ascribing molecular or genetic functions such as enzyme, protein or tRNA to the given sequences. Transcriptional activity imparts the ability to adapt to temporal fluctuations in an organism's environment. The structure and abundance of transcripts within the transcriptome reflect this necessity by means of temporal differential expression and alternative splicing of transcripts. Transcriptomes are thus inherently dynamic, a moment in time providing just a snapshot of the transcriptional potential for an organism.

In the absence of a genome for *B. braunii*, global transcriptomics by next generation RNA sequencing (RNA-Seq) provides a highly sensitive and accurate replacement in the field of gene function and biological process discovery in non-model organisms. RNA-Seq using NGS technology can identify full sets of transcripts, including rare and/or novel unannotated, alternatively spliced and fusion transcripts, with a much larger range in detection of expression and greater sequencing depth of 100–1000 reads per bp of a transcript than the traditional Sanger sequencing (Martin & Wang, 2011; Ozsolak & Milos, 2010). Furthermore, RNA-Seq circumvents the issue of bacterial contamination because the concentration of polyadenylated mRNAs can be enriched during sample preparation since prokaryotic mRNAs are not polyadenylated.

Without a reference genome, transcriptomic sequence data must be assembled into contiguous sequence (contigs) *de novo*- to achieve sequences that are the full-length of the original transcripts or as close as possible. The necessity for assembly is due to the short sequencing reads of Next Generation Sequencing (NGS) platforms of 35-500 bp (Metzker, 2009).

De novo transcriptome assembly remains a challenging and relatively new process. Existing tools for genomic assembly are not equipped to handle the complexity inherent to transcriptome data. A primary reason for this difficulty is that genome assemblers identify areas of repetition in genomes based on the number of reads covering that area - an assumption that would nullify any information on differential expression amongst transcripts from the dataset and render all abundant transcripts as repetitive. Furthermore, splice variants from the same gene add to the complexity, as they share exons and are therefore difficult to resolve. Lastly, transcriptome data may be strand specific - assemblers must cater for this possibility, allowing resolution of over-lapping sense and anti-sense transcripts (Johnson *et al*., 2012; Merino *et al*., 1994).

This chapter aims to purify, sequence, annotate and archive the full *B. braunii* transcriptome by incorporation of mRNA sequence data from a 28-hour time series of cultures in two different light-regimes. Analyses in subsequent chapters will focus on temporal and photo-regime effects on the transcriptional profile of *B. braunii*.

In this chapter, the process from *B. braunii* cell cultures through RNA purification and sequencing to annotated transcriptome is described. The acquired bioinformatic data are then analyzed and ascribed EC numbers, Gene Ontology terms and KEGG pathway identifiers constituting a database that can be published and made available for future research.

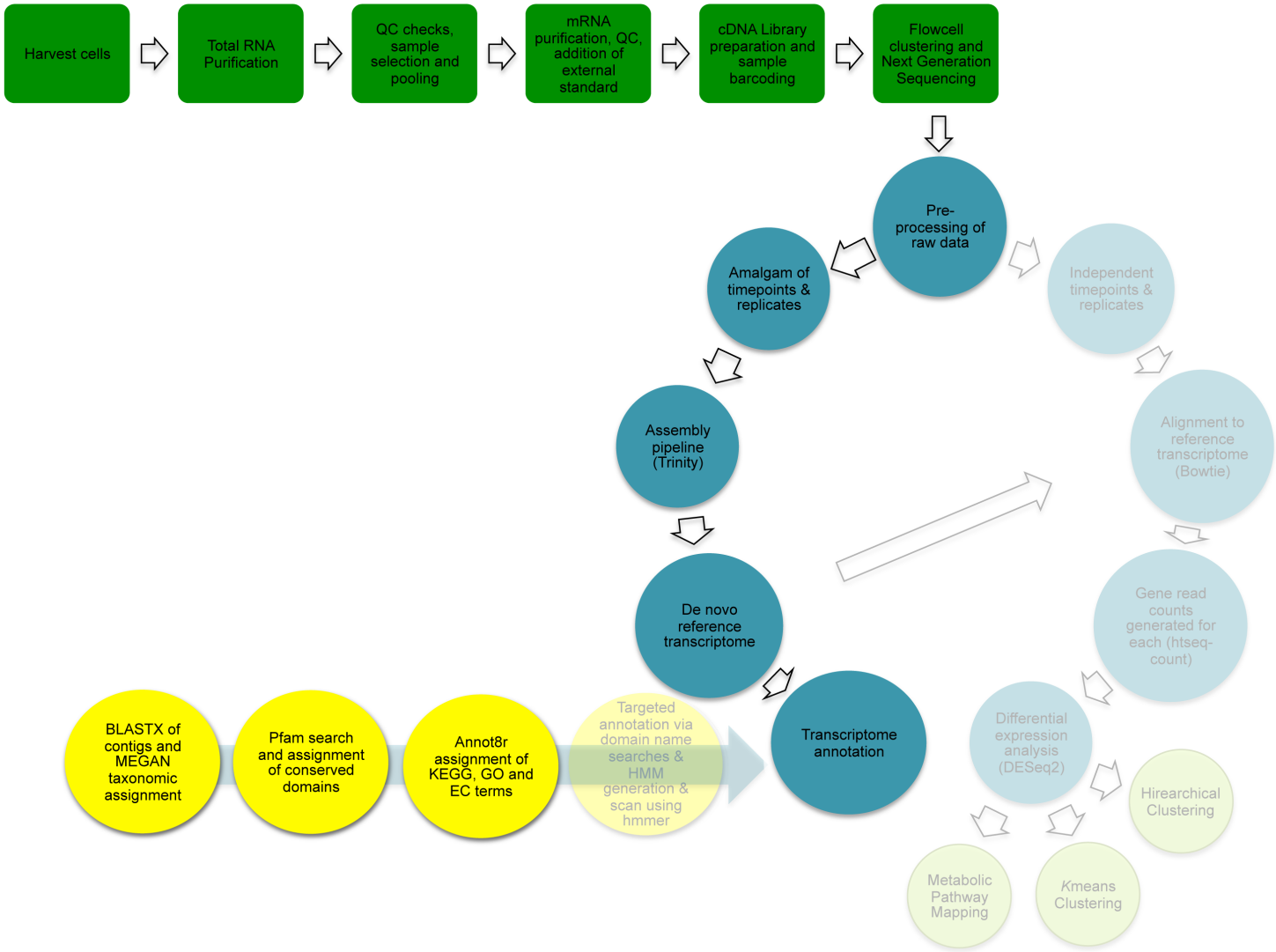


Figure 1 Workflow showing experimental and bioinformatic procedures

Practical laboratory methods (green) through to bioinformatic analysis (teal blue) used to investigate the transcriptomics of *B. braunii* over a 28-hour time series under two different photo-regimes. Chapter 3 describes the generation and annotation of the *Botryococcus braunii* transcriptome (green, teal blue and yellow) sequence. Faded aspects are addressed in Chapters 4 and 5.

3.2 Materials and Methods

3.2.1 Harvest and preparation of samples

192 cultures (50 ml) of *B. braunii* were inoculated at approximately 0.2 OD_{680} and grown according to conditions specified in General Materials and Methods, section *2.4*, with exception of the photoperiod; in this case the algae was provided with 12 hours of illumination in a 24 hour diel cycle (LD) or constant light (LL). After five days 12 cultures were harvested per timepoint, as explained in General Materials and Methods section *2.5,* at eight four-hourly intervals, from 08:00 until 12:00 the following day Following this total RNA was purified, quantified and quality checked as described in General Materials and Methods sections *2.8* and *2.11.2.1* respectively.

The standardised measure of RNA quality, RNA Integrity Number (RIN) (Schroeder *et al*., 2006), was used to summarise the results of microfluidic quality assessment of RNA by the Agilent Bioanalyzer. RIN is computed based on nine critical features in the RNA electropherogram profile, each with its own integrator such as peak area, intensity or ratio, and collectively the ratios of these characteristics determine the RIN and thereby RNA quality.

Time-points were replicated (n=3) by allocating nine samples with the highest RIN into three pools, such that each sample contributed 2 μg RNA and each pool had a total of 6 μg RNA. Concentrations of the pools were checked again according to the protocol detailed in General Materials and Methods section *2.11.2.1*. External RNA Control Consortion (ERCC) ExFold RNA Spike- In Control Mix (Ambion Life Technologies) was added to a volume of each sample pool containing 5 μg RNA before proceeding with mRNA purification as detailed in General Materials and Methods section *2.10*.

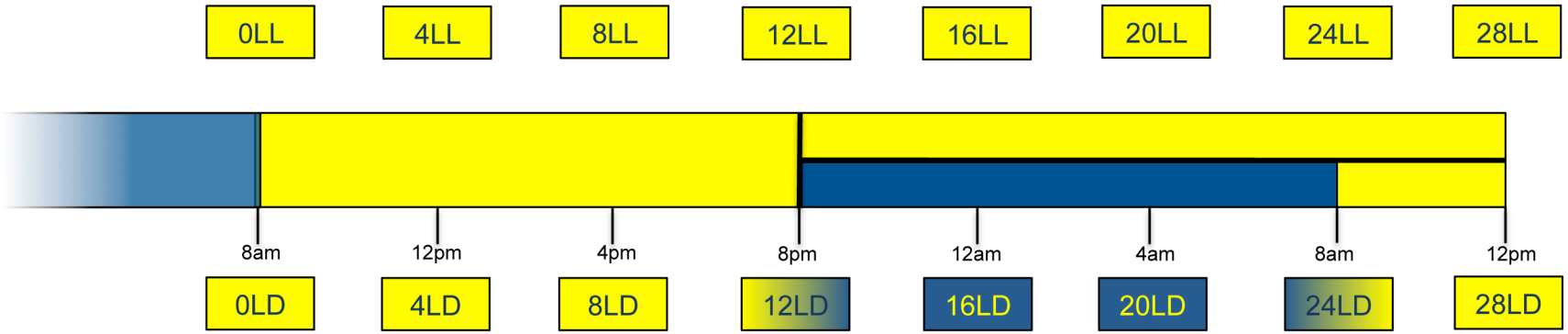


Figure 2 Sample timeline

A timeline showing the actual time at which time-points were taken. The photoperiodic samples beneath the timeline were those in 12 hours light and 12 hours of dark. Yellow boxes indicate samples were taken during a light phase (or LL). Blue boxes indicate samples taken during a dark phase. LL samples above the timeline were those under continuous light conditions, therefore all boxes are yellow.

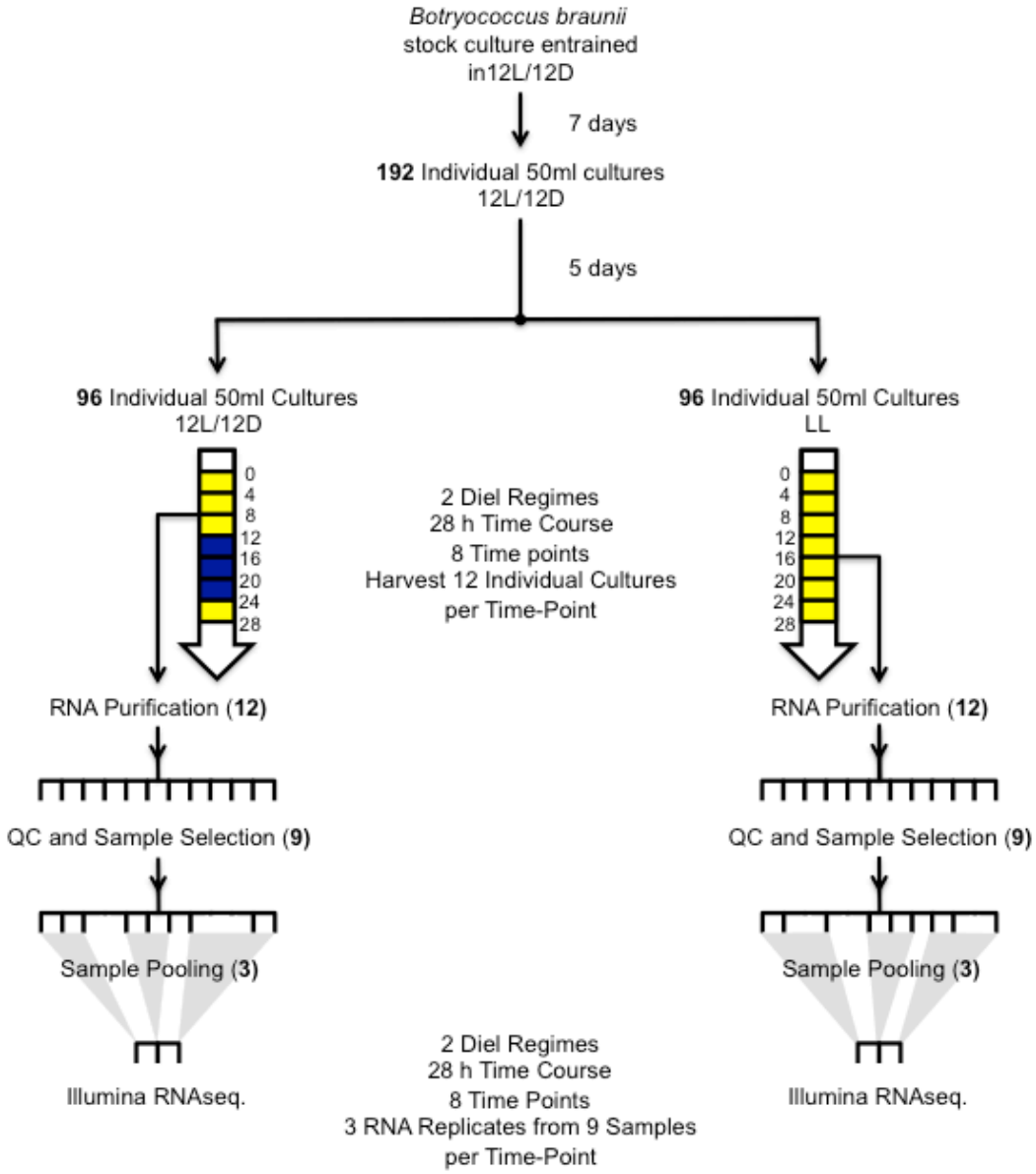


Figure 3 Schematic of sample processing
Diagram showing the processes from *B.braunii* sample harvesting to prepared Illumina cDNA libraries.

3.2.2 Construction of cDNA libraries and sequencing

cDNA libraries were prepared and amplified using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre) and Kapa HiFi Library Amplification Kit (Kapa Biosystems). Briefly, RNA was fragmented before cDNA was synthesised from the antisense strand. cDNA fragments were terminal tagged at both 3' and 5' ends and the library was purified. During the subsequent amplification by polymerase chain reaction, the libraries were barcoded for multiplexing (*i.e.* where multiple samples are loaded per sequencing lane) by replacement of the reverse primer with 1 μl 10 μM Epicentre Index Primer, ensuring that each library within one sequencing lane was identifiable by a different primer. Primers were selected such that at each position of the index code to be read during sequencing, there was at least one A or C and at least one T or G; ensuring accuracy of signal detection and therefore confident delineation between indexed samples. (Primer sequences are given in Table 1 in the Appendix). Finally, cDNA fragment libraries were purified using Agencourt AMpure XP system (Beckman Coulter). Library concentration and size distribution of fragments were determined using an Agilent Bioanalyzer and Expert Software version 1.2, before libraries were diluted to 10 nmol and stored at -20°C.

For sequencing, the libraries were distributed into eight multiplexed groups and each group loaded into one lane of an Illumina 8-lane flow cell at a concentration of 6.5 pmol. Using the TruSeq PE Cluster Kit v3 - cBOT (Illumina), the single cDNA molecules available within each library were covalently bound by complementary adaptors to the flow cell surface and clonal clusters generated using the Illumina cBOT automated system. The flow cell was finally prepared for paired-end sequencing using TruSeq SBS Kit HS v3 (Illumina) and run on an Illumina HiSeq 2000 by the University of Exeter Sequencing Service.

3.2.3 Pre-processing of raw reads and sequence assembly

Raw Illumina short read FASTQ formatted data were imported into the RobiNA software application (Lohse *et al*., 2012) for pre-processing quality checks, adaptor sequence removal, trimming and filtering. Modules were selected to assess base call quality, consecutive homopolymers, k-mer and base call frequencies, over-represented sequences and basic statistics. Base call quality or PHRED score (Figure 4) was calculated by $Q_{PHRED} = -10*\log(P_e)$ where P is the probability of error in the assignment of a nucleotide to a specific position within a read, computed by the sequencing platform.

Reads originating from adaptor sequences and bases with a Q_{PHRED} below 20 were removed using the standard RobiNA module set and the defaults associated (Lohse *et al.*, 2012).

Q_{PHRED}	Probability of base call error	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Figure 4 Illumina quality scoring
Q_{PHRED} scores and their corresponding probabilities of error in the base call made in that read position and the inferred accuracy of each call.

The FASTQ/A trimmer tool, FASTX Toolkit version 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/download.html) was used to remove the first 15 bases from each read after initial quality checks indicated this as an area of sequence bias **Figure 7** (a)). Reads with missing paired ends after trimming and filtering (orphans) were removed from the dataset using a BioPython (Cock *et al*., 2009) script available from GitHub (https://github.com/lexnederbragt/denovo-assembly-tutorial/blob/master/scripts/interleave_pairs.py) as part of the BioPython toolkit (version 1.64-1.fc19.x86_64).

The processed sequence reads from all replicates at all time-points were then compiled together into 2 files, one of forward reads and another of reverse reads.

The reference transcriptome assembly was then generated using the Trinity *de novo* assembly pipeline (Grabherr *et al*., 2011) with the compiled forward and reverse reads of all samples as input. 20Gb of system memory and 26 central processing units (CPUs) were allocated to the assembly process, all other parameters remained as default. The Trinity method of assembly is based on De Bruijn graphs. At every possible position the sequencing reads were broken into 25 nucleotide-long sub-sequences called k-mers (or 25-mers) such that each k-mer overlapped with the last by K-1. Edges on a De Bruijn graph were generated from each k-mer and Trinity attempted to create a single path between nodes, each node generated from a K-1, using the k-mers as bridges if they created an overlap between nodes. Reads overlapping in sequence generated from an mRNA transcript were thus appended together resolving the longest possible single path between the nodes, via the edges created from k-mers. In this way stretches of contiguous sequence (contigs) were computed from reads generated from fragmented transcripts. Alternatively spliced transcript information was retained. In a first round of assembly, the Trinity software reported only the parts of alternatively spliced transcripts that were unique then pooled together sequences that overlapped and finally used the generated K-1mers to construct full length transcripts where there were reads supporting junctions between the unique contigs (Grabherr *et al*., 2011).

Trinity software outputs the sequence assembly as a FASTA file of nucleotide sequence, delineated into transcripts. The assembled transcripts were translated into predicted Open Reading Frames (ORFs) using a Python script (http://atgc-tools.googlecode.com/files/seqs_processor_and_translator_bin_V136_AGCT.py).

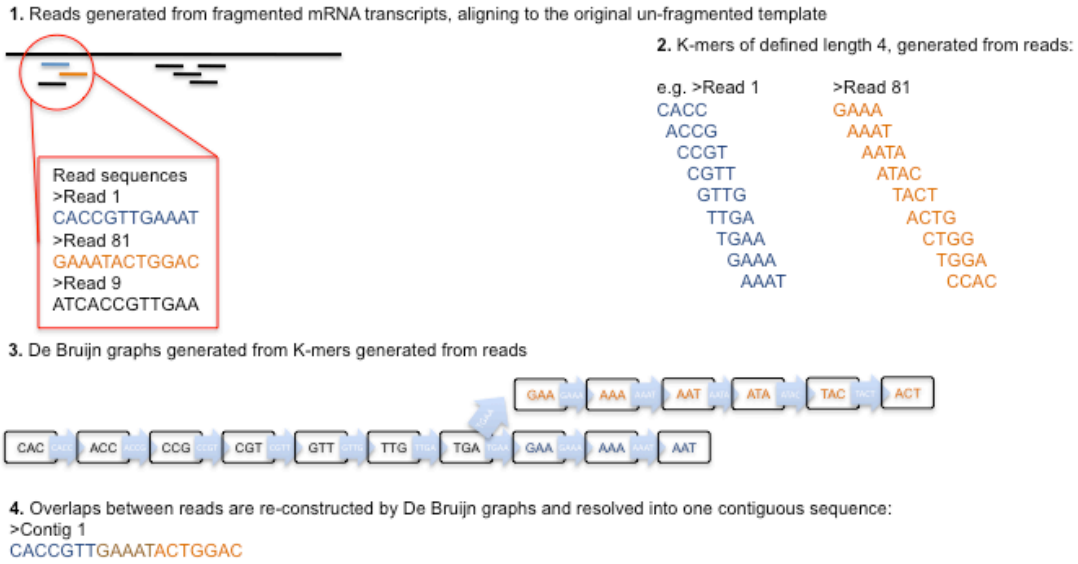


Figure 5 Generation of contigs by De Bruijn graphs

Reads aligning to the mRNA transcript (long black line) they originated from (1.). K-mers are generated from each read (colour-coded in blue and orange) (2.) and contiguous sequence identical to the original mRNA transcript is reconstructed by the overlap between Reads 1 and 2 (3.)

3.2.4 Functional annotation

Methods of gene annotation use different attributes of sequences to infer function, for example the NCBI Basic Local Alignment Search Tool (BLAST) uses overall sequence homology to infer the function and phylogeny of a sequence, whereas the use of the protein database, Pfam, will only take into account conserved domains and characterise a sequence based on the function of its conserved domains. Annotation of the transcriptome was performed using several different approaches. Methods for annotation including sequence homology by BLAST, assignment of Enzyme Commission numbers (EC), Kyoto Encyclopedia of Genes and Genomes identifiers (KEGG) and Gene Ontology (GO) terms and conserved domain architecture are described below.

3.2.4.1 BLAST sequence homology

The BLAST algorithm generates an Expect Value (E value), which indicates whether the BLAST hit to an unknown nucleotide or protein sequence query can be expected by chance from within the specified database of known sequences. The E value is exponentially inversely proportionate to the score of an alignment along the entirety of two nucleotide or protein sequences (protein in this case) so that the closer to 0 an E value is, the more confidence can be assigned to match. The E value is computed from the bit score, which is a normalised transformation of the raw score and also takes into account sequence length and database size.

The raw score is given by $R = aI + bX - cO - dG$, where I is the number of identities in the alignment, a is the ascribed integer for each identity, b is the ascribed integer for each mismatch and X is the number of mismatched nucleotides or amino acids. c is the penalty for opening a gap and the number of gaps is given by O. d is the penalty for each "-" (where a potential identity may have been) in the gap and G is the size of the gap given by the total number of "-" characters. By default $a = 1$, b = -3, c = 5, d = 2.

The bit score is given by $S = \lambda R - \ln K / \ln 2$ where λ and K are constants equal to 1.37 and 0.7111 respectively and n is the length of the sequence database queried.

Finally the E value is given by $E = mn2^{-S}$ where m is the length of the unknown sequence query.

Using tools available as part of the BLAST+ toolkit (Camacho *et al*., 2009) a local, standalone search of the downloaded NCBI Non-Redundant Protein Database (Pruitt *et al*., 2013) was performed using the translated nucleotide sequences of all transcripts in the *B. braunii* transcriptome as a query with an E value cutoff of $1e^{-8}$ and the number of result descriptions and alignments displayed in the output restricted to

three; eight CPUs were used. The results of the BLAST search were parsed to produce a table format of the top hits for each ORF using a custom Perl script, blast_parse_pw2tab.pl (courtesy of University of Exeter Microbial Biofuels Bioinformatics group).

Both the query nucleotide sequence and the database nucleotide sequences were translated in all six reading frames, effectively performing a protein-to-protein search but with enhanced sensitivity. Translated nucleotide sequences were used because this method circumvents the issue of frame shift and other ambiguities that may cause failure to detect certain open reading frames in other programs.

3.2.4.2 Gene Ontology, KEGG and EC features - Annot8r

GO is a controlled vocabulary developed and curated specifically to provide a consistent way of describing gene products and to allow collaborating databases to be searched using uniform terminology (Gene Ontology Consortium, 2004). KEGG is a database of KEGG Orthologies (KOs), which are the basis of enzyme and metabolic pathway maps, modules and functional hierarchies that is curated to be applicable to all organisms. KOs are manually defined for all proteins that correspond to nodes within KEGG pathways, hierarchies or modules so that once a sequence is annotated with a KO identifier, the sequence can be characterised in terms of metabolic function (Kanehisa *et al*., 2013). EC numbers are assigned to gene products based upon their enzymatic function, indicating the chemical reaction catalyzed and the recommended nomenclature for enzymes performing that function. Different enzymes with the same chemical or catalytic function will be assigned the same EC number. EC numbers therefore encode a general classification system in which the first digit of an EC number indicates the broadest functional category of an enzyme and the last digit, the most specific functional category (Hu *et al*., 2012).

Using the Annot8r (version 1.1.1) platform (Schmid & Blaxter, 2008) to search the *B. braunii* transcriptome predicted ORFs for sequence homology against databases downloaded from EBI (Uniprot) (Binns *et al*., 2009), Expasy (Gasteiger, 2003), KEGG (Kanehisa *et al*., 2013) and GO, GO terms, EC numbers and KO annotations were assigned to gene products identified in the *B. braunii* transcriptome. Annot8r is a tool for the annotation of non-model organisms based on their similarity to annotated sequences and uses databases from which uninformative entries have been removed. A cutoff level of acceptable similarity based on a BLAST bit score of 55 was specified and the 10 BLAST hits with highest score were used for each annotation.

3.2.4.3 Protein family assignment - Pfam search

The predicted Open Reading Frames (ORFs) of the *B. braunii* transcriptome were searched against the PfamA Hidden Markov Model database using the command line tool hmmsearch from HMMER version 3.1b1 (Eddy, 2011) to identify conserved and detectable domains of the *B. braunii* transcriptome.

Protein families were identified based on conserved domains by the generation of models incorporating parameters of position-specific probabilities of variation in amino acids, insertions and deletions into a Multiple Segment Viterbi (MSV) algorithm (Eddy, 2011).

3.2.5 Removal of contaminating sequences

Part way through the BLAST search of the *B. braunii* predicted open reading frames, the distribution of BLAST annotations was checked. The BLAST hits resulting from the search of the *B. braunii* transcriptome were assigned to specific taxa using the metagenomics analysis software, *MEtaGenome* ANalyzer (MEGAN). A large number of transcripts aligning with sequences from the buff-tailed bumblebee species, *Bombus terrestris,* were identified contaminating the dataset. Contamination was traced to another project within the department where the operator had not decontaminated equipment following use. These contaminating sequences were removed from the assembly using the standalone Expressed Sequence Tag (EST) alignment program, gmap, which aligns cDNA sequences to a genome. An index was created from a FASTA format file of the genome of *B. terrestris* (downloaded from NCBI refseq database, accession PRJNA68545) using the gmap_build utility with the output parameter –f specifying a PSL format output, and assembled *B. braunii* transcripts were aligned using the gmap program, version 2013-08-19 (Wu & Watanabe, 2005). Those transcripts that successfully aligned were removed from the gmap output file.

B. terrestris sequences were not removed from the unassembled read data retained in individual time-points and replicates used for later differential expression analyses. Sequences assembled into transcripts could be removed with a higher degree of confidence that they originated from *B. terrestris* and were not legitimate *B. braunii* sequences. Furthermore, once transcripts assembled from reads from *B. terrestris* templates were removed from the reference transcriptome dataset, read data originating from *B. terrestris* within individual samples would not successfully align to the reference and thus will be automatically discounted from differential expression analysis described in subsequent chapters. The annotation pipeline, with exception of the BLAST search, was repeated after *B. terrestris* sequence removal.

3.3 Results

3.3.1 Preparation of samples, construction of cDNA libraries and sequencing

Sufficient high quality RNA (RNA Integrity Number (RIN) >= 7) was purified for preparation of 47 cDNA libraries (Table 1). An insufficient quantity of algal sample was available for the preparation of a third replicate library for time-point 12LD due to unplanned exposure to white light during harvest as the green filter fell from a laboratory light.

The standardised measure of RNA quality, RIN (Schroeder *et al*., 2006), was used to summarise the results of microfluidic quality assessment of RNA by the Agilent Bioanalyzer. In all samples selected for cDNA library synthesis, 25S and 18S ribosomal subunits were automatically resolved by the analysis software and are clearly distinguishable above the baseline, which is flat for the most part but in some cases becomes slightly jagged in the inter-region between 18S and 25S peaks corresponding to low level degradation (Figure 6) Additional peaks immediately flanking the left of the 18S peak correspond to smaller chloroplast 16S ribosomal RNA; cytosolic 8S and cytosolic and mitochondrial 5S subunit peaks were also visible in order from right to left between the 16S peak and the lower marker in most traces, with a varying degree of degradation of these small subunits.

Sample name	Quality (RIN)	Total quantity RNA (μg)
0D P1	7.80	5.66
P2	8.10	6.84
P3	7.80	4.44
4D P1	7.97	6.00
P2	7.77	3.83
P3	8.00	6.43
8D P1	7.87	4.01
P2	7.90	5.14
P3	7.83	6.51
12D P1	7.27	6.38
P2	7.23	6.60
16D P1	7.20	4.65
P2	7.43	4.11
P3	7.47	4.88
20D P1	7.50	7.47
P2	7.70	2.93
P3	7.70	5.60
24D P1	7.67	7.33
P2	7.60	3.59
P3	7.73	4.05
28D P1	7.23	6.41
P2	7.23	5.49
P3	7.43	4.04
0C P1	7.47	4.83
P2	7.50	4.06
P3	7.33	3.62
4C P1	6.97	4.54
P2	7.03	5.60
P3	7.17	5.81
8C P1	7.50	5.18
P2	7.53	4.92
P3	7.50	3.59
12C P1	7.60	3.95
P2	7.63	5.63
P3	7.57	5.39
16C P1	7.77	4.30
P2	7.77	2.85
P3	7.70	9.69
20C P1	7.30	3.59
P2	7.27	4.48
P3	7.20	4.76
24C P1	7.27	5.07
P2	7.27	4.56
P3	7.20	2.35
28C P1	7.13	4.39
P2	7.23	5.06
P3	7.17	4.49

Table 1 RNA quality and concentration by sample

Sample names are given with the RNA quality metric, RNA Integrity Number, and the total mass of RNA within each pooled sample.

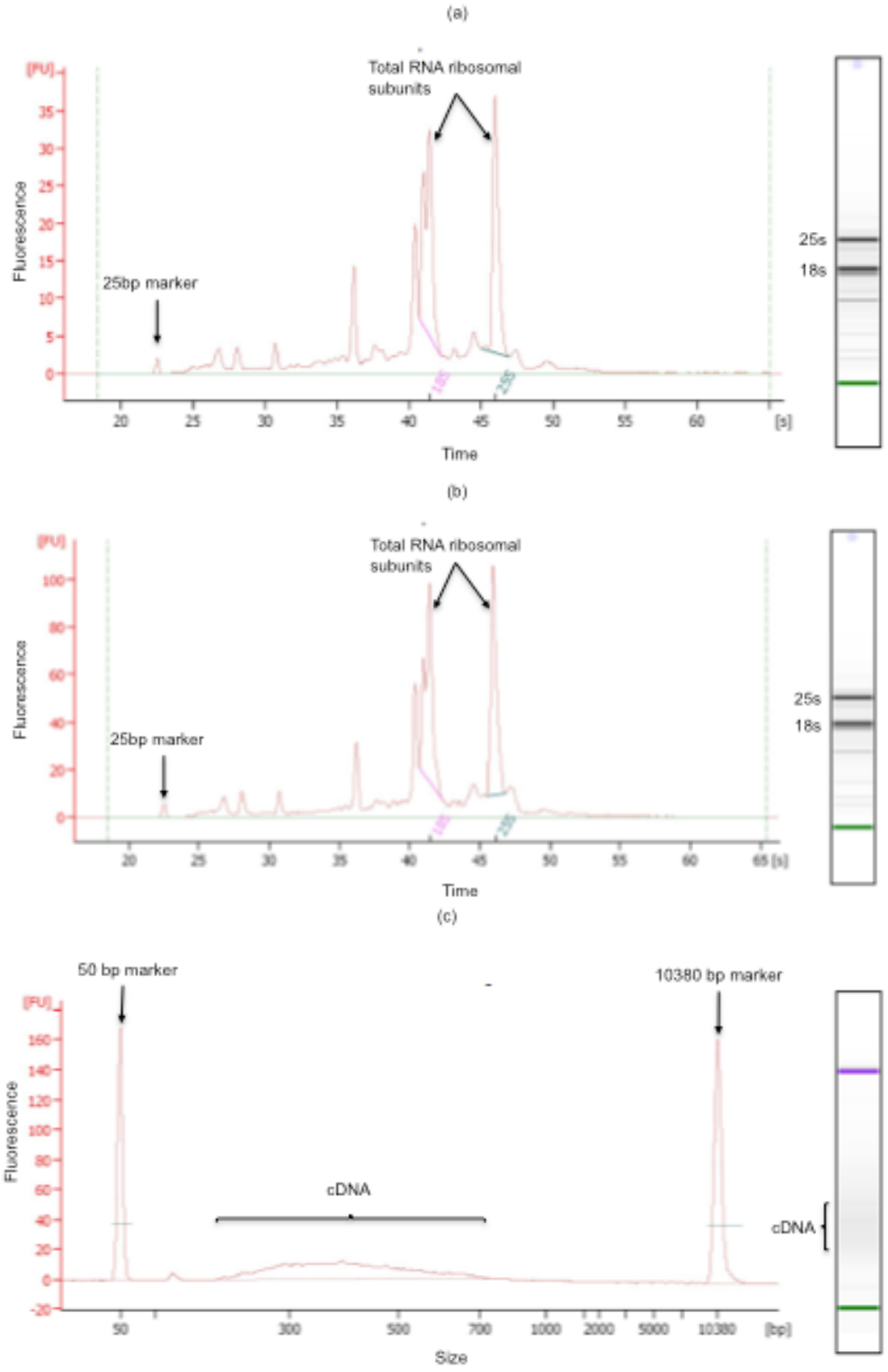


Figure 6 Bioanalyzer trace images

An individual Total RNA sample that comprised one of the three samples to make up a library 4DP1 (a). Library 4DP1 trace image, showing little degradation after samples had been freeze-thawed for pooling together (b) and (c) the final resulting cDNA library. As machine read- outs it was not possible to improve.quality of trace images.

3.3.2 Pre-processing of raw reads and sequence assembly

One run of an 8-lane flow cell in an Illumina HiSeq 2000 generated 201.1 GB of sequence data in the form of 2,011,057,142 100 bp paired-end reads, with an average quality score (Q_{PHRED}) of 31.79, 80% of which had a score above 30. Quality scores and read numbers varied little between lanes of the flow cell (Table 2).

Read numbers were broadly comparable between samples, with the exception of samples 0LDP2 (8,574,544 reads), 20LDP2 (5,410 reads) and 24LDP2 (15,990 reads), which had read numbers that were tenfold, ten-thousandfold and one-thousandfold lower than the rest. The sample with the highest read number was 4DP2 (93,954,594 reads) and the lowest, aside from the above mentioned low yielding samples, was 12DP2 (11,304,852 reads).

Sample by lane	Reads	Mean Q_{PHRED} per sample and per lane (in bold)
Lane 1:	**274384682**	**31.44**
0DP1	36247102	32.8
12DP1	43011070	31.8
16DP1	38389266	31.12
20DP1	52085516	31.88
4DP1	47597684	30.83
8DP1	57054044	30.17
Lane 2:	**201309482**	**31.52**
0DP2	8574544	31.03
12DP2	11304852	31.55
16DP2	57728966	31.88
20DP2	5410	31.99
4DP2	93954594	31.74
8DP2	29741116	30.94
Lane 3:	**328671420**	**30.88**
0DP3	66008024	31.35
16DP3	54624678	31.72
20DP3	51131726	31.24
24DP3	47178738	31.2
4DP3	57142384	28.77
8DP3	52585870	30.99
Lane 4:	**252950656**	**32.19**
0CP1	53292828	31.32
0CP2	36808954	32.51
24DP1	29415474	32.94
28DP1	52039832	31.89
28DP3	42199832	31.84
4CP1	39193736	32.65
Lane 5:	**212198780**	**30.56**
0CP3	41519896	31.25
12CP1	36863208	29.98
24DP2	15990	30.51
28DP2	46756868	29.47
4CP2	41508958	30.96
8CP1	45533860	31.18
Lane 6:	**249669922**	**32.66**
12CP2	49160850	32.75
16CP1	39820990	32.49
20CP1	62438664	32.84
24CP1	32763602	32.52
28CP1	35751318	32.22
8CP2	29734498	33.12
Lane 7:	**254138558**	**32.72**
12CP3	51103922	32.36
16CP2	48037602	32.67
20CP2	41272646	33.07
24CP2	35067082	31.99
28CP2	31646894	33.08
8CP3	47010412	33.17
Lane 8:	**237733642**	**32.52**
16CP3	50901334	33
20CP3	38700076	32.82
24CP3	54196174	32.39
28CP3	42404120	33.02
4CP3	51531938	31.35
Total	**2,011,057,142**	**31.79**

Table 2 Sequence reads by lane

Multiplexed cDNA library groups shown with corresponding lane number, total reads per lane and quality of each, mean quality per lane is in bold.

FastQC analysis (

Figure 7) of raw sequences was performed on all samples but represented in this chapter by the report from forward reads of 0DP1, which is typical of the whole dataset. Mean per-base quality decreased below Q_{PHRED} 32 beyond position 85 of the reads. In the first 13 positions of the reads there was bias in the proportion of nucleotides, with cytosine (C) content being disproportionately high (~ 31%) between positions 2 and 5 but falling to represent less than 20% of sequence content from positions 5 to 13. However, C was less represented than other nucleotides across the length of the reads. Thymine (T) was under-represented from positions 1 to 7 (≤ 20%) but remained at the expected 25% throughout the rest of the read length. Guanine (G) representation differentiated less significantly from an even ratio between the four bases than C and T but was somewhat higher between positions 6 and 9. The adenosine (A) ratio remained balanced at ~ 25% for the entire read length, although was slightly more variable between positions 1 and 13 than for the rest of the read length.

Trimming and filtering removed the areas of sequence content ratio imbalance and lower quality sequences (Figure 8). FastQC reports for all other samples after trimming and filtering are available in the Appendix, Figures 7- 12). 1,476,587,564 cleaned reads remained after the cleaning and filtering steps, 73% of the total sequenced reads.

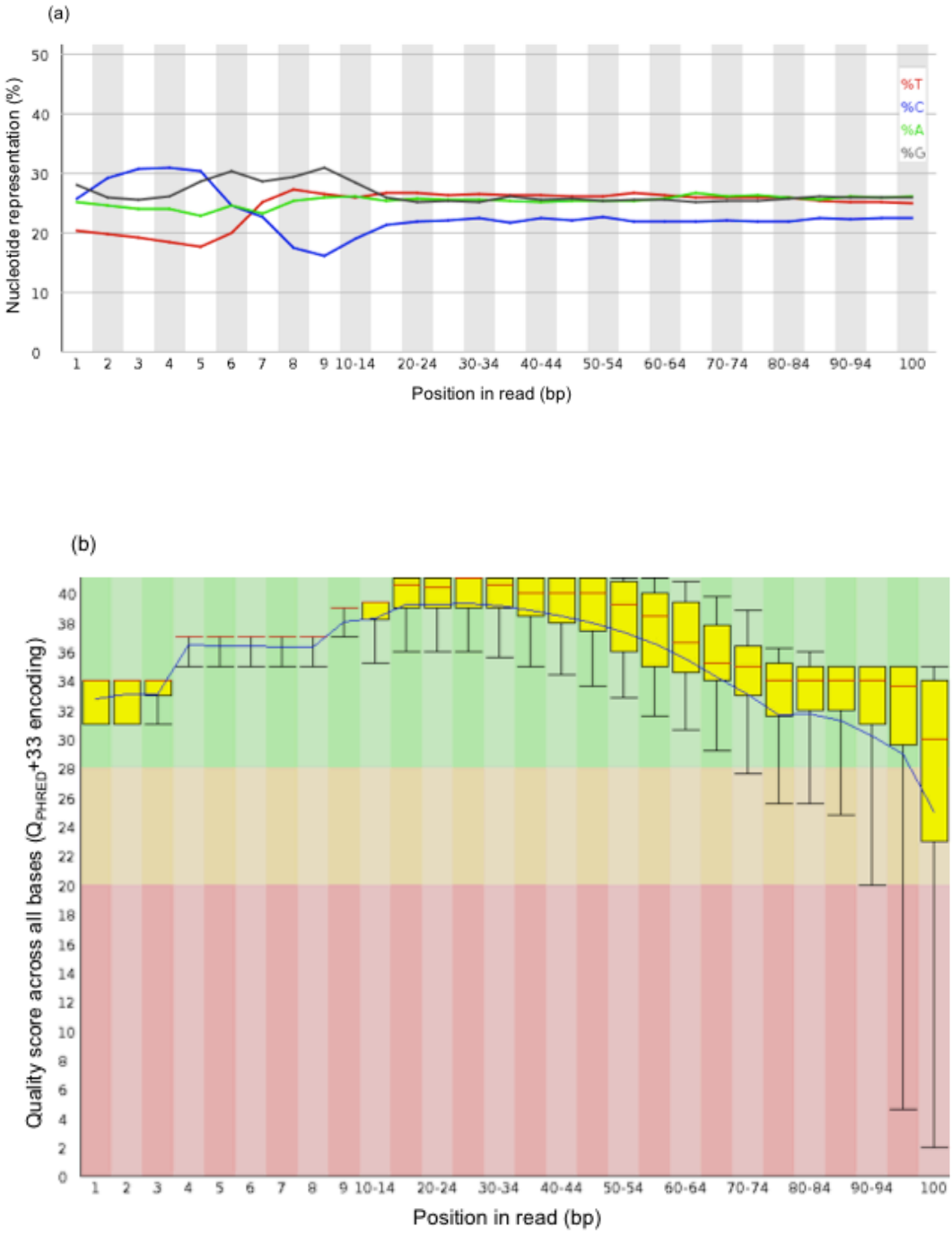


Figure 7 Raw read FastQC quality checks

FastQC report of forward reads from Sample 0DP1 showing positional ratio of bases across read length (a) and a box-whisker plot of per-base sequence quality (b) before cleaning and trimming. In (b) mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

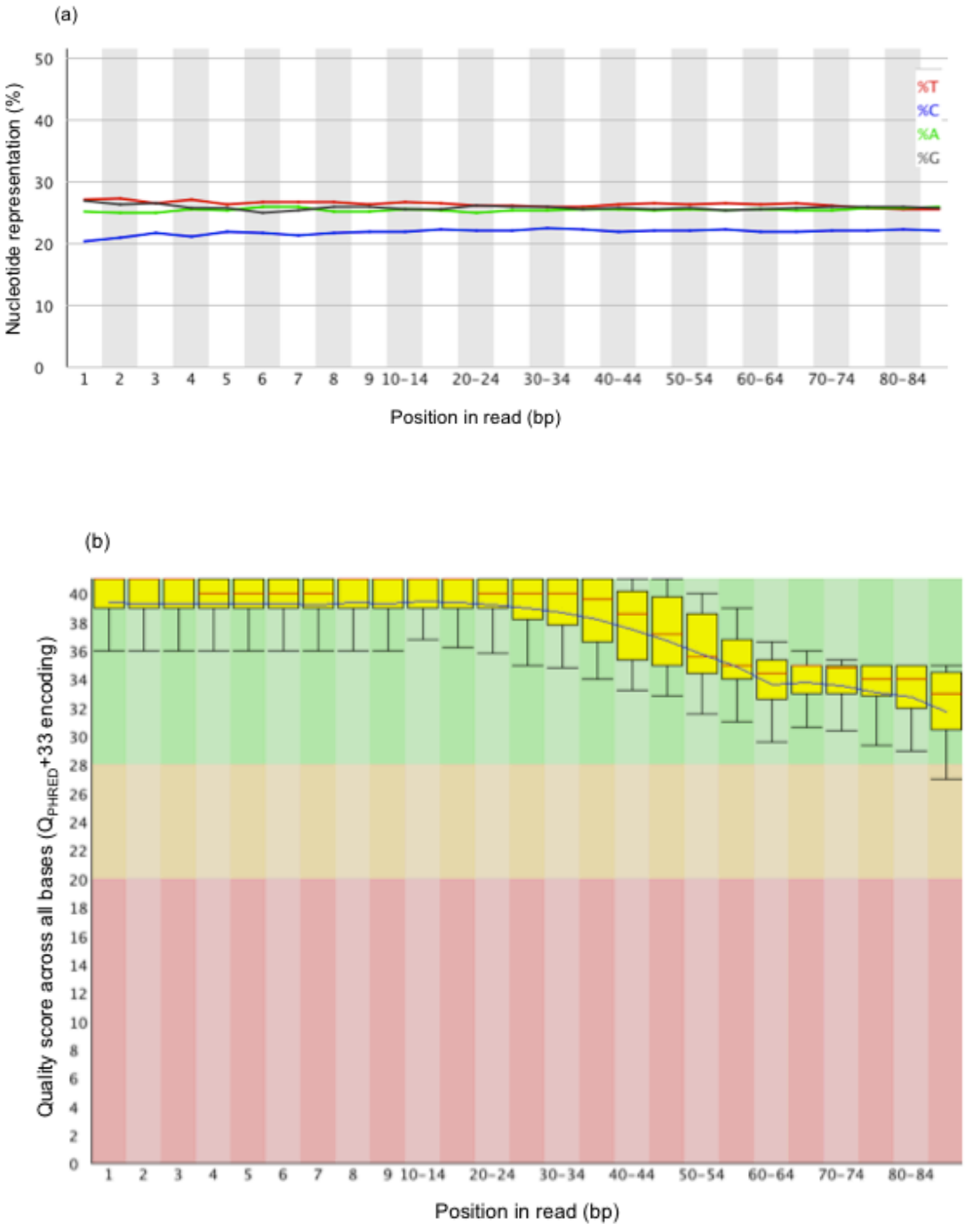


Figure 8 Processed read FastQC quality checks

FastQC report of forward reads from Sample 0DP1 showing positional ratio of bases across read length (a) and a box-whisker plot of per-base sequence quality (b) before cleaning and trimming. In (b) mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

After stringent trimming and filtering steps, paired-end reads were assembled into 331,569 transcripts in an assembly of 336 million nucleotides (nt) in length.

Putative *B. braunii* transcripts were generated in the form of transcripts by the Trinity pipeline using the De Bruijn graph algorithm.

The GC content of the resulting assembly was 48%. 47% of transcripts were 500 nt or longer. *B. braunii* transcriptome transcript length distribution was plotted (Figure 9) using a Python script (http://wiki.bioinformatics.ucdavis.edu/index.php/Count_fasta.pl) with a specified interval size of 50. The most numerous transcript size was 200-250 nt with nearly 60,000 transcripts. Nearly 40,000 and 30,000 transcripts were contained within the 250-300 and 300-350 nt size bands respectively. Size categories of 500-550 nt long or less were all represented by less than 10,000 transcripts, the frequency getting progressively smaller with increasing length. The assembly N50 transcript length was 2,089 nt, indicating that 50% of the entire assembly is represented by transcripts of this length or lower. The longest transcript was 34,166 nt and the shortest 201 nt (Table 3).

After the standard quality control steps, 27,930 transcripts discovered via analysis in following chapters and confirmed by sequence homology to originate from the bumblebee, *B. terrestris,* were removed before resuming downstream analysis. 303,639 transcripts remained within the assembly.

Category	Total
Total transcripts (#)	331,569
Total assembly length (nt)	336,339,021
Transcripts > 500 nt (#)	155,786
Transcripts > 1K nt (#)	90,579
Transcripts > 10K nt (#)	696
Median transcript size (nt)	462
Mean transcript size (nt)	1014
Longest transcript (nt)	34,166
Shortest transcript (nt)	201
GC content (%)	47.47
N50 transcript length (nt)	2,089

Table 3 Assembly statistics

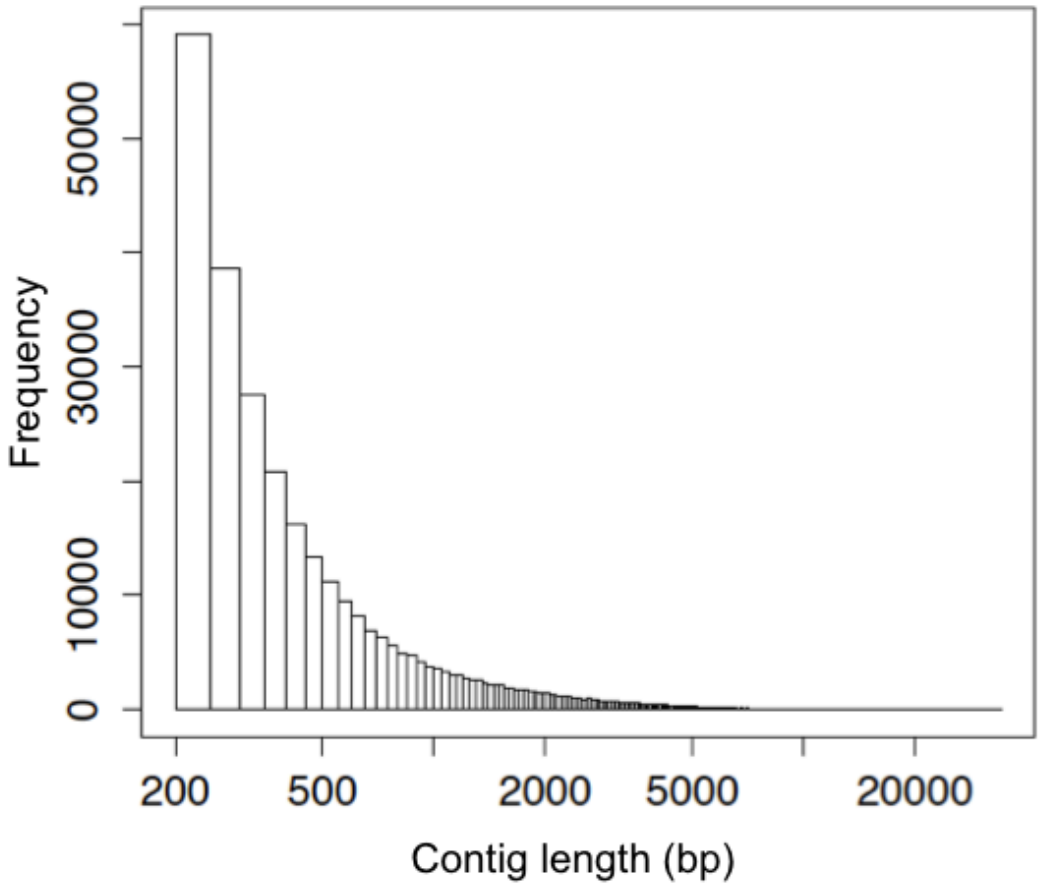


Figure 9 Contig length distribution
The length distribution of *B. braunii* transcriptome assembly contigs, measured in base pairs.

3.3.3 Functional annotation of assembled B. braunii *transcriptome*

3.3.3.1 BLAST search of NCBI Non-Redundant Protein Database

Of 331,569 total transcripts, 128,800 had a significant hit (E value ≤ 1^{-08}) in the NCBI Non-Redundant protein database, representing 39% of the total *B. braunii* transcriptome. Of the 128,800 transcripts with significant hits, 35,677 were assigned to the *Viridiplantae*.

The BLAST annotation data were imported into MEGAN software version 5.4.2 (Huson *et al*., 2011) which assesses the taxonomic content based on the Lowest Common Ancestor (LCA) using NCBI taxonomy. The taxonomic distribution of BLAST hits for all 331,569 transcripts was visualised as a phylogenetic tree generated from summarised read counts (Figure 10). Of the transcripts assigned to taxa within the *Viridiplantae*, the green algae represented 82% (29,271 transcripts). 24,606 transcripts were assigned to the Trebouxiophycae class and orders below. 5,865 transcripts were assigned to the Streptophyta and most of these to the Embryophyta (5,858). Just 670 transcripts were assigned to the Bryophytes, all of which had BLAST hits with *Physomitrella patens.* The remaining 4,978 of Streptophyta transcripts were assigned under Tracheophyta, which is the taxonomic group that encompasses all vascular land plants.

A large number of transcripts (19,086) were annotated under Insecta and 13,280 of these transcripts showed sequence homology with the genus, *Bombus* (bumblebees). Other large nodes on the graph were those depicting 21,418 transcripts assigned to bacteria, 18,811 transcripts assigned to Saccharomycetes and 10,278 to the Amoebozoa. 202,771 transcripts remained un- assigned to taxa by MEGAN and 5,445 transcripts had no BLAST hits above the cutoff, amounting to 61% of the assembly.

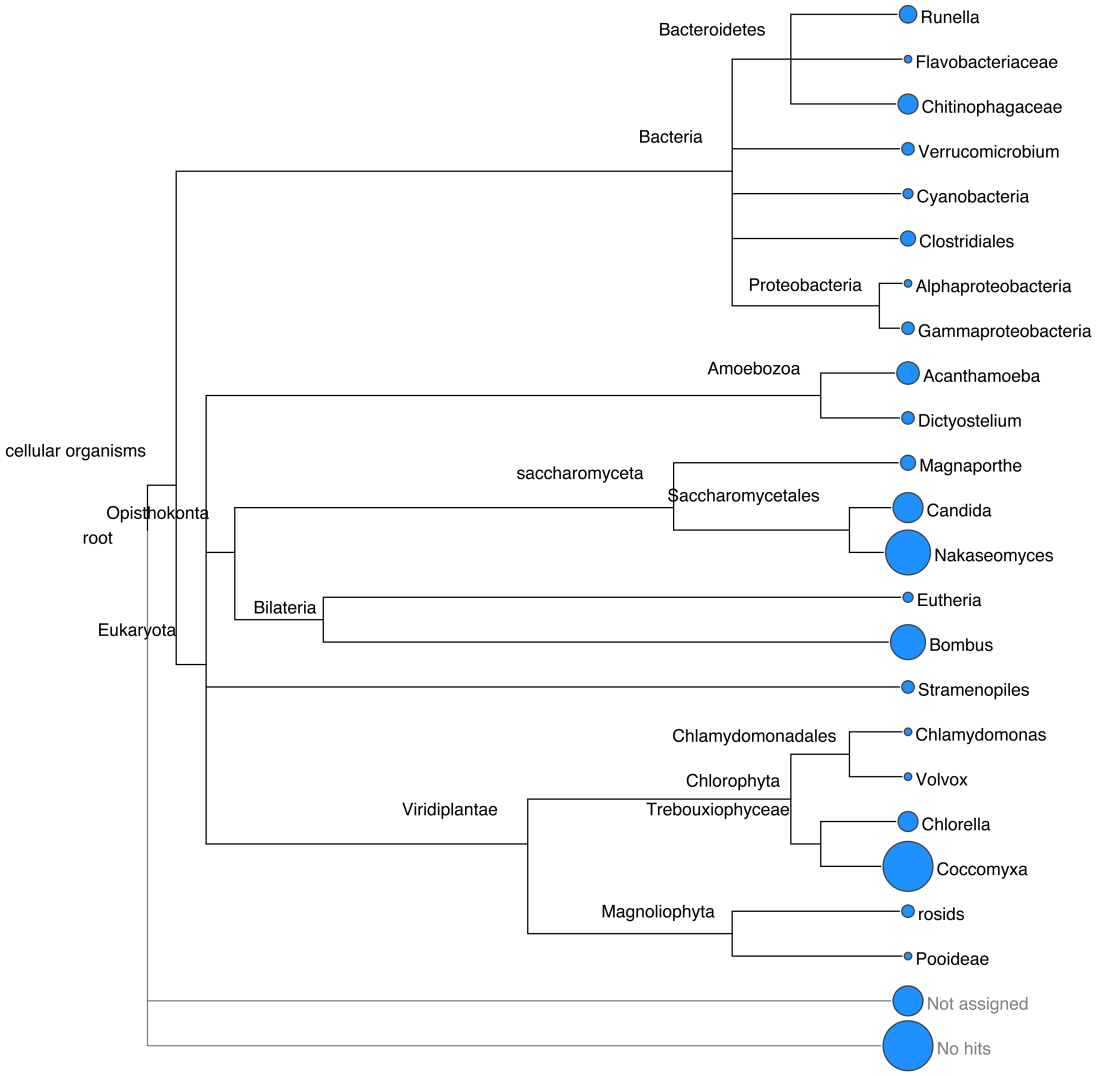


Figure 10 Taxonomic distribution of BLAST annotations

BLAST annotated contigs grouped by taxa represented by blue nodes, which are scaled logarithmically to represent the summarised number of contigs within each taxa. The annotations were filtered to show nodes only for taxa supported by 100 contigs or more.

3.3.3.2 Gene Ontology annotation

Once GO terms had been assigned to as many predicted ORFs as possible in the transcriptome, a summarised representation of the functional attributes of the annotations was determined using GOSlim annotation, generated by CateGOrizer. The summary was created by considering only the parent GO term to a gene product in cases where both parent and child terms had been assigned, *e.g.* using only the term for protein kinase activity where both protein kinase activity and protein phosphorylation activity were assigned because protein kinase (child term) is involved in protein phosphorylation (parent term).

Terms assigned to *B. braunii* predicted ORFs were segregated into Biological Process, Cellular Component and Molecular Function categories, which segregate terms in a GOSlim annotation according to the root of ontology using GOroot (Table 4). 16,442 annotated predicted ORFs could be mapped by single count to one of the three GO_ROOT ancestor terms. Single counts indicated that whilst there may be multiple pathways from a child term through its ancestor terms to the root, these have been consolidated to only one that best represents them all. Only 1,560 counts of the transcriptome gene products were designated within Cellular Component, the smallest number assigned to a category. The category containing the largest number of single counts (10,516) of the annotated gene products of the *B. braunii* transcriptome was Biological Process, then Molecular Function with the second most numerous single counts (4,366).

GO_SLIM2 annotations were generated with the terms within each ontology root category to show the lower level activity distribution amongst gene products within the transcriptome.

Category	Counts
Biological Process	10,516
Molecular Function	4,366
Cellular Component	1,560

Table 4 Assignment of read counts to GO categories

21,817 transcripts were annotated under the Biological Process root category - the most numerous of the three (Figure 11). Within the biological process subcategory, metabolism accounted for the largest number of terms assigned, with 3,619 counts. Development (1,754 counts), cell organisation and biogenesis (1,421 counts) were all also abundantly represented as independent subcategories within the biological component root category. Transcripts related to precursor metabolite and energy generation, cell growth and cytoplasm biogenesis were amongst the lowly represented subcategories with only 63, 47 and 10 counts respectively.

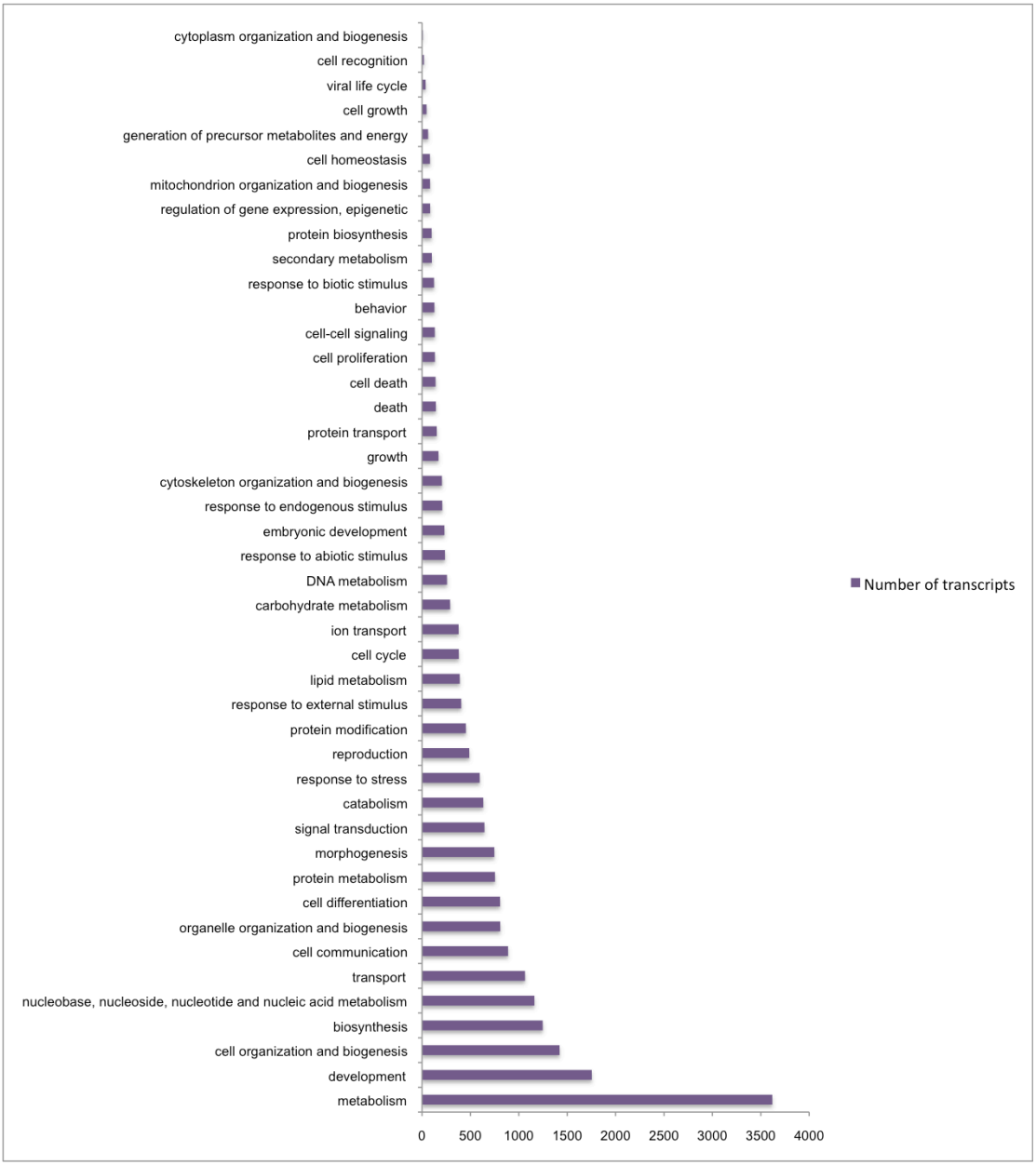


Figure 11 Distribution of Biological Process GOs
The number of *B. braunii* transcriptome contigs annotated with GOs in subcategories describing biological processes.

Under Molecular Function, 7,624 transcript ontologies were assigned and association with catalytic (2,710 counts), transferase (881 counts) and generic binding (872 counts) activities were most abundant (Figure 12). Kinase and protein kinase GOs were well assigned with 215 and 78 counts respectively. Signal transducer, receptor activity and binding, enzyme regulator and DNA binding activities were all in the mid-range of abundance and protein kinase activity towards the lower end of this group of counts, ranging from 125 to 78. Nutrient reservoir activity and calcium ion binding were notably low, the lowest of all representation across all root categories with one count each.

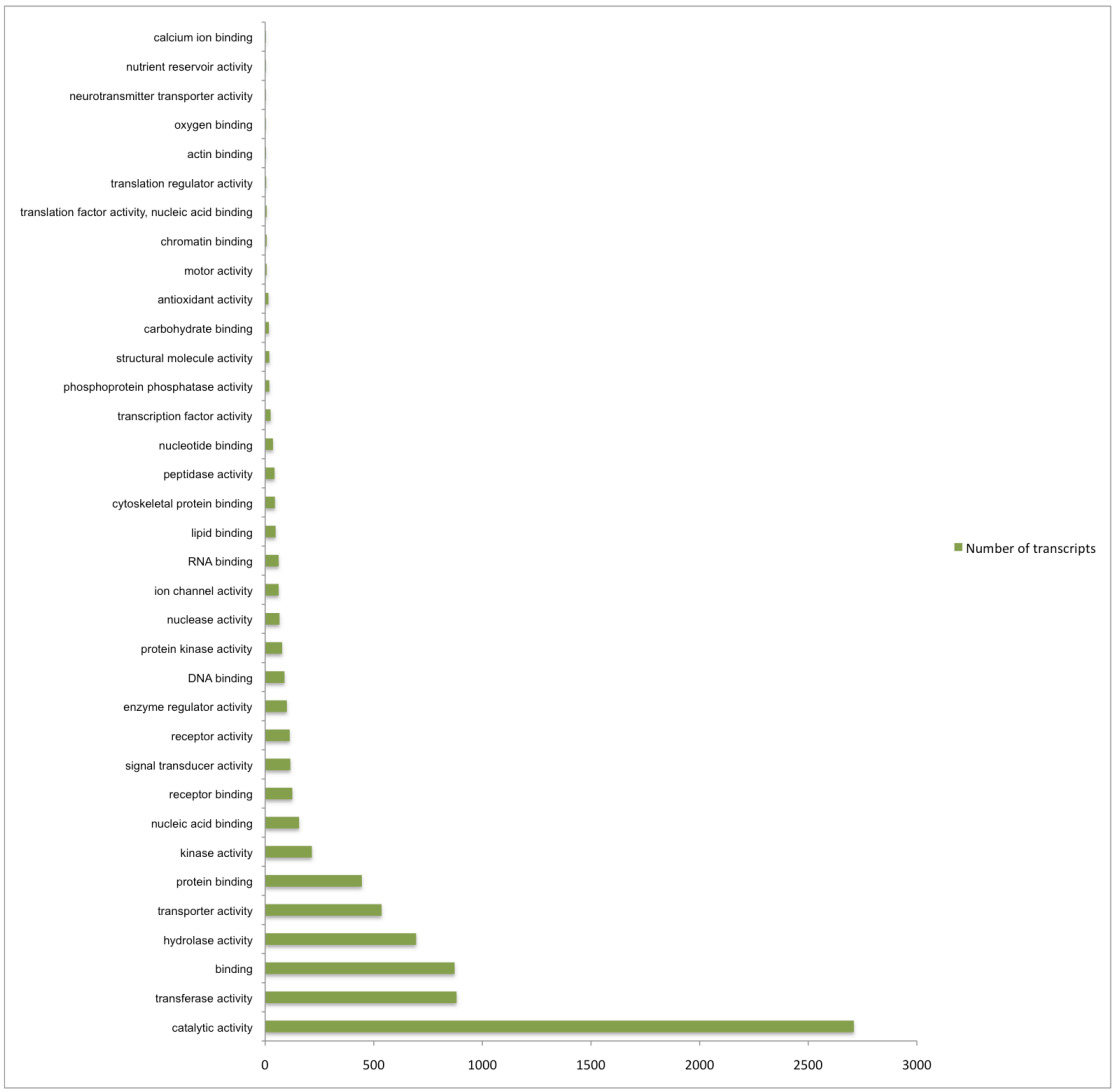


Figure 12 Distribution of Molecular Function GOs
The number of *B. braunii* transcriptome contigs annotated with GOs in subcategories describing molecular functions.

4,472 transcripts were annotated with GO terms within the Cellular Component root category (Figure 13). After intracellular (1,123 counts), cytoplasm (543) and nucleus (289) were the highest represented. Cytoplasmic membrane-bound vesicle, Golgi apparatus and extracellular matrix all featured in the mid-range of counts with 77, 26 and 24 respectively. Notably, lipid particle GO-annotated transcripts were few, only three counts.

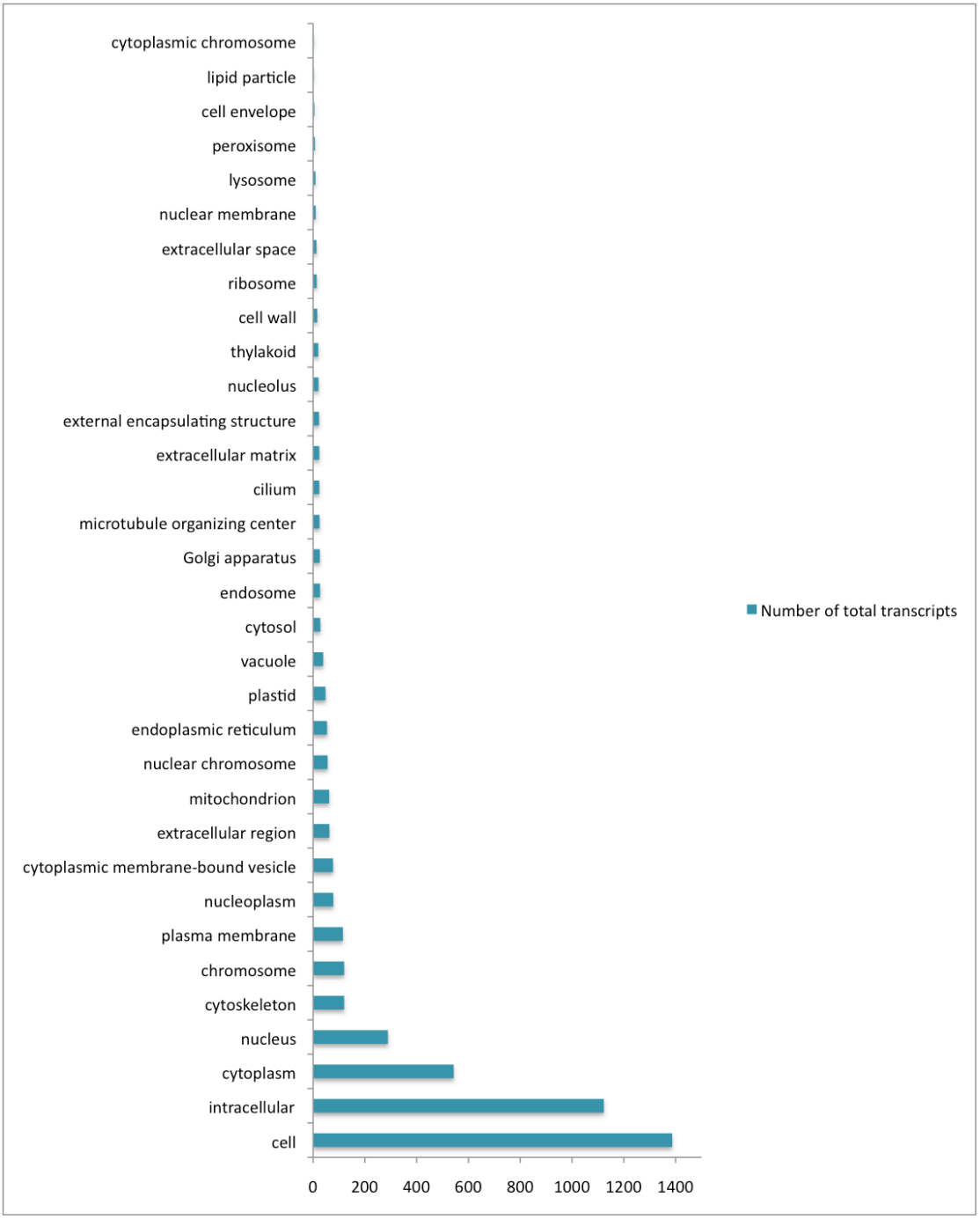


Figure 13 Distribution of Cellular Component GOs
The number of *B. braunii* transcriptome contigs annotated with GOs in subcategories describing cellular processes.

3.3.3.3 KEGG pathway analysis

Using BLAST annotation data and MEGAN software, transcripts that were annotated as *Viridiplantae* were extracted from the BLAST annotations of the whole *B. braunii* transcriptome assembly. A KEGG pathway map of the *Viridiplantae* transcripts was generated (Figure 14) using Annot8r assigned KOs and the KEGG Search and Color tool, which maps transcripts annotated with either EC numbers or KOs to model pathways within its own database. In the following text, the 5-digit KEGG pathway map identifiers are stated in brackets. 1,047 transcripts were mapped to the KEGG Global Metabolic Pathway map (01100), 396 of these encompassed within the overview map, Biosynthesis of Secondary Metabolites (01110).

Notable areas with sets of pathways lacking coverage by *B. braunii* transcripts on the Global Metabolic Pathway map were “Glycan Biosynthesis and Metabolism”, “Xenobiotics Biodegradation and Metabolism”, Metabolism of Cofactors and Vitamins”.

70 gene products were identified from within the Cell Cycle pathway, although 21 were missing. 30 of the Eukaryotic basal transcription factors were identified and just six missing.

The KEGG group subdivisions of “Carbohydrate Metabolism and Energy Metabolism” were well covered, such as the citrate cycle (00020), with 26 gene products identified and just five missing enzymes from the pathway (phosphoenolpyruvate carboxykinase; pyruvate ferredoxin oxidoreductase; alpha subunit; isocitrate--homoisocitrate dehydrogenase; 2-oxoglutarate ferredoxin oxidoreductase subunit alpha; succinyl-CoA:acetate CoA-transferase and malate dehydrogenase (quinone). 95 gene products of the oxidative phosphorylation pathway (00190) were mapped and 32 protein subunits and enzymes of the Photosynthesis pathway (00195) were identified, although a further 30 remained unmapped. Lipid Metabolism is also well mapped with 46 gene products identified involved in Fatty Acid Metabolism (01212) from pathways localised to the cytoplasm or plastid and mitochondria and endoplasmic reticulum. However, gaps in some Lipid Metabolism pathways remain, for example in steroid biosynthesis, with 14 missing components and 23 mapped.

Discounting the overview pathways (Biosynthesis of Secondary Metabolites, Carbon Metabolism, Fatty Acid Metabolism) purine metabolism (00230) had the highest number of transcripts ascribed with 126 counts, ribosome (03010) was next with 122 counts, followed by biosynthesis of amino acids (01230, 122 counts). Amongst the others were carbon metabolism (01200), spliceosome (03040), RNA transport (03013), endocytosis (04144), MAPK signaling (04010) and RNA degradation (03018) (Figure 14).

26 transcripts were assigned KOs from within the Carbon Fixation pathway in photosynthetic organisms (00710), leaving only four unmapped. 23 transcripts mapped to the terpenoid backbone synthesis pathway (00900) enzymes, completing the 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway (MEP/DOXP pathway) route into squalene synthesis (Figure 15). In the following text EC numbers are stated in brackets. MEP/DOXP enzymes identified in the dataset were 1-deoxy-D-xylulose-5-phosphate synthase (EC:2.2.1.7), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (EC:1.1.1.267), 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (EC:2.7.7.60), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC:2.7.1.148), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC:4.6.1.12), (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (EC:1.17.7.1), 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (EC:1.17.1.2) and isopentenyl-diphosphate delta-isomerase (EC:5.3.3.2). Squalene synthase was also mapped using KO assignment to transcripts.

Unmapped edges in the whole metabolome map occur whereby enzymes are present to synthesise the nodes either side but not convert between the two nodes, creating a gap on the map.

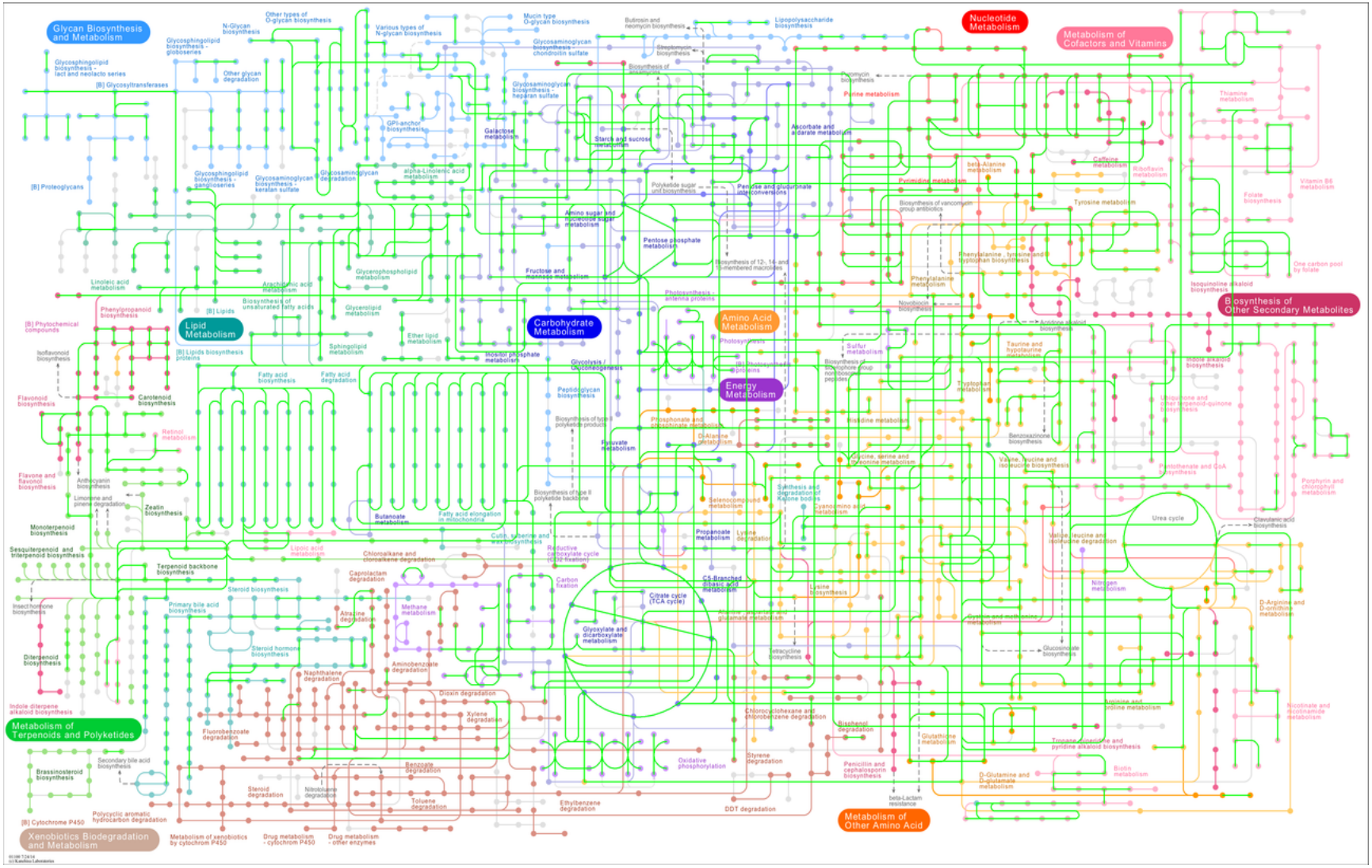


Figure 14 KEGG pathways map of *B. braunii* transcriptome

Transcriptome contigs filtered to only those assigned within *Viridiplantae* and mapped using KEGG Orthologies to gene products (green lines) in the KEGG reference pathway. Compounds are represented by nodes. Unmapped products are colour coded by KEGG according to pathway.

Top 20 KEGG pathways	Number of transcripts mapped to pathway
Purine metabolism	126
Ribosome	122
Biosynthesis of amino acids	122
Carbon metabolism	114
Spliceosome	108
Pyrimidine metabolism	96
Oxidative phosphorylation	95
Protein processing in endoplasmic reticulum	89
RNA transport	87
Ubiquitin mediated proteolysis	81
Huntington's disease	78
Endocytosis	76
Epstein-Barr virus infection	75
Cell cycle	70
Pathways in cancer	67
Cell cycle - yeast	66
Parkinson's disease	66
MAPK signaling pathway	64
Alzheimer's disease	62
ABC transporters	62
RNA degradation	60

Table 5 Top 20 KEGG Pathways

The KEGG pathways with the annotated *B. braunii* transcripts mapped to them. Transcripts were mapped using KEGG Orthology identifiers.

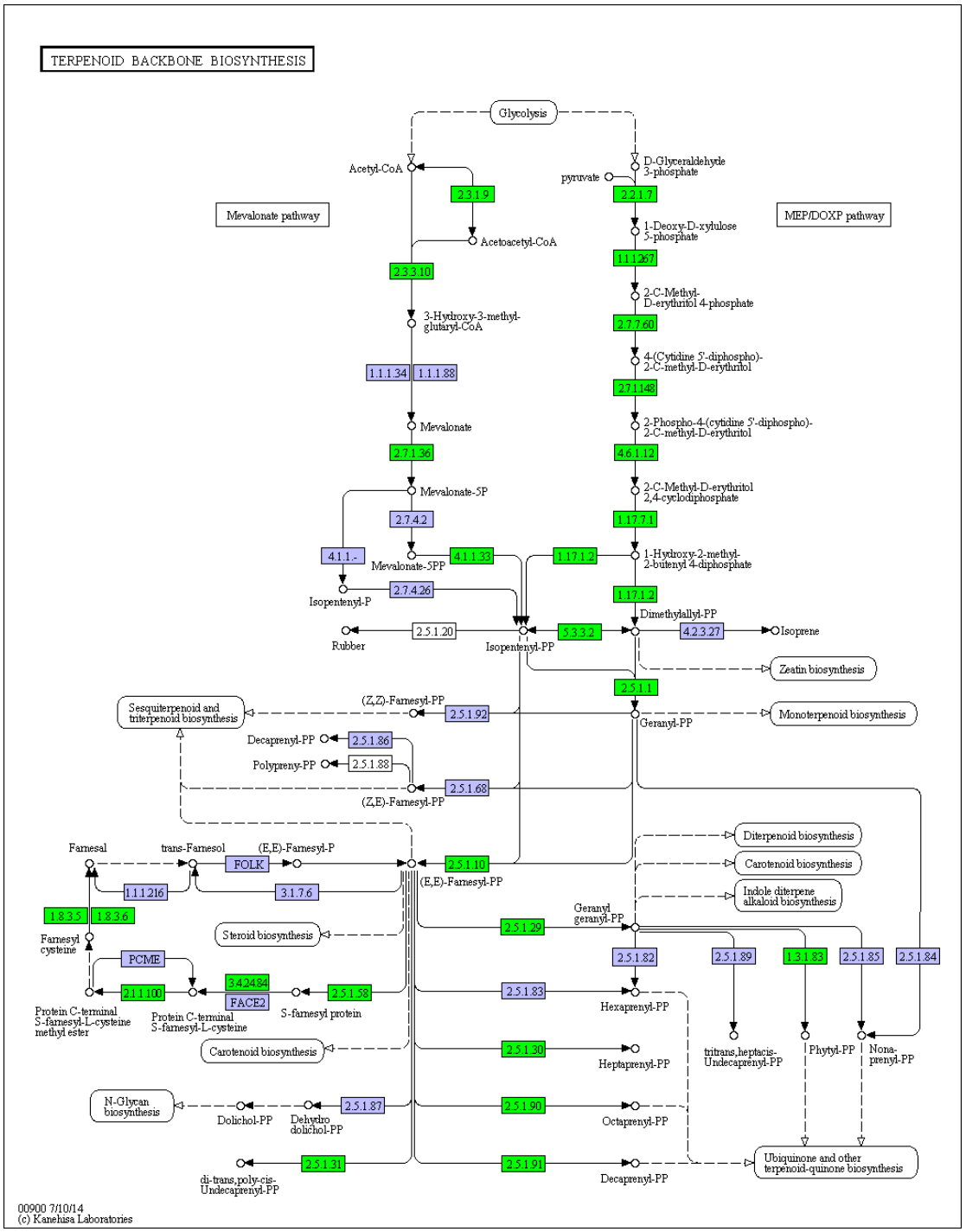


Figure 15 KEGG terpenoid backbone pathway

Transcriptome contigs filtered to only those with Viridiplantae BLAST annotation and mapped with EC numbers to the enzymes (green boxes) in the KEGG terpenoid backbone reference pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

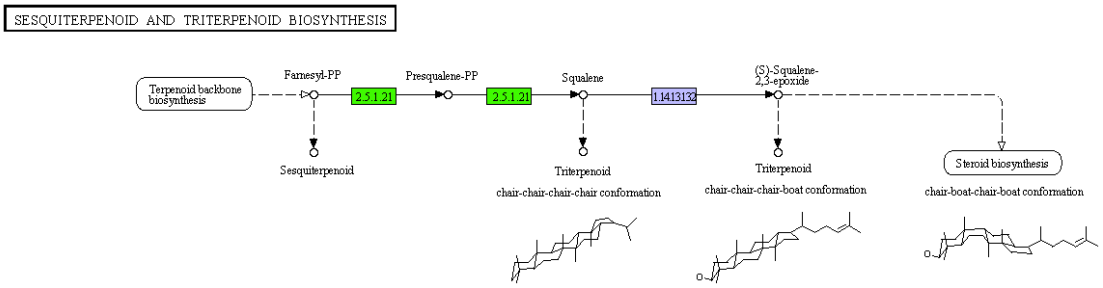


Figure 16 KEGG squalene synthesis pathway

Transcriptome contigs filtered to only those with *Viridiplantae* BLAST annotations and mapped with EC numbers to enzymes (green boxes) in the squalene synthesis pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs from squalene synthesis to other pathways are shown in rounded-edge boxes. Circles represent compounds.

3.3.3.4 Assignment of protein domain families using Pfam scan

273,949 transcripts were assigned Pfam A family domains, accession numbers and E values for use as a custom database for targeted pathways and transcript annotation analysis in forthcoming chapters. E values were assigned based on the alignment score over the total length of the alignment and a separate E value calculated from the score only in the regions of conserved domains within the alignment. 87,837 transcripts had an E value equal to or less than $1e^{-5}$ along the length of domain alignments. 97,484 transcripts had an E value equal to or less than $1e^{-5}$ along the length of the alignment. The most numerous three Pfam domains assigned to transcripts in the *B. braunii* transcriptome with domain alignment scores of $E \leq 1e^{-5}$ were the protein kinase (Pkinase; PF00069; 2,645 transcripts), protein tyrosine-protein kinase (Pkinase_Tyr; PF07714; 2,386 transcripts) and WD domain (WD40; PF00400; 760) transcripts. Both protein kinase domain categories were substantially over-represented in comparison to any other, with an approximate threefold difference in number of transcripts assigned with protein kinase domains and those with the next most numerous; WD domain. The Eukaryotic export protein superfamily, ABC transporters, feature twice in the top 20 most represented Pfam domains in the *B. braunii* transcriptome, 318 ABC type II transmembrane domain (ABC2-membrane; PF01061) and 590 ABC ATP-binding domain (ABC_tran; PF00005) annotated transcripts respectively.

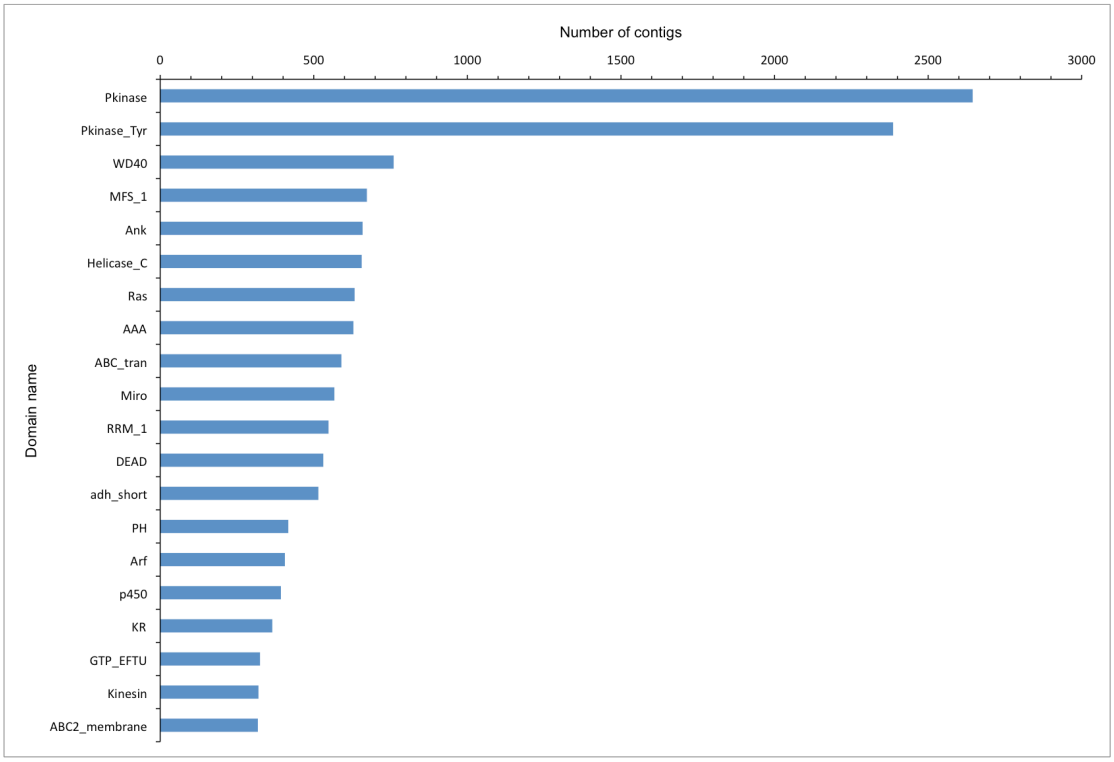


Figure 17 Top 20 Pfam domains in *B. braunii* transcriptome

The top 20 most numerous Pfam domains assigned to contigs in the *B. braunii* transcriptome with an E value cutoff of $1e^{-5}$ scored from the conserved domain regions of alignments.

3.4 Discussion

3.4.1 Pre-processing of raw reads and sequence assembly

A high quality profile of *B. braunii* race B, Guadeloupe strain expression encompassing the transcriptional repertoire in two different photoperiodic regimes over the full daily cycle, was generated. Samples were harvested during the exponential growth phase when gene expression and enzyme activity involved in the production of botryococcenes are maximal (Matsushima *et al*., 2012; *et al*., 2000). A comprehensive database of annotated transcripts is presented, replicated to a higher degree (n=3, compared to n=1), encompassing more conditions (eight time-points under two conditions, compared to seven time-points under one condition) and yielding a larger and more contiguous transcriptome than those already available (303,639 transcripts with mean length 1,014, compared to 46,422 and 756 respectively) (Guarnieri *et al*., 2011; Molnár *et al*., 2012; Yang *et al*., 2013).

Information pertaining to different isoforms of each transcript was retained using the Trinity assembly pipeline, so not only can sequence transcripts be annotated but insights into gene splicing events in *B. braunii* are available.

The technical reliability of RNA-Seq technology was capitalised upon by maximising biological replication over technical; n = 3 for all samples contributing to the transcriptome, except for D12, for which only two replicates exist due to insufficient high integrity RNA. Furthermore, each biological replicate is comprised of three RNA samples, extracted separately from independent cultures from the same time-point. In such a way, each *bona fide* biological replicate is supported by three pseudo-replicates, mitigating the effects of individual anomalies. However, there is no way of detecting these anomalies or delineating between the pseudo-replicates.

Quality checking of raw reads prior to sequencing revealed imbalances in the nucleotide composition between positions 1 and 13 of reads Figure 7 (a) and Figures 7- 12 in Appendix for all sample quality reports). Such observations are common in Illumina sequencing and are due to the use of random hexamers as primers during cDNA synthesis. Non-uniformity in the location of reads aligning along the length of a transcript has been attributed to the bias induced by random hexamer primers, although the inclusion of an RNA fragmentation step prior to priming for cDNA synthesis mitigates against this effect and has been implemented in the *B. braunii* library preparation for this work. In some library preparation protocols Oligo(dT) primers are used as an alternative to random hexamers but are substantially biased toward the 3’ end of mRNAs because of the poly-A enrichment complementary to the poly-T primer (Hansen *et al*., 2010).

Lower quality of sequence such as that observed from position 85 onwards in the *B. braunii* raw reads (**Figure 7** (b) and Figure 7- 12 in Appendix for all sample quality reports) is also characteristic of Illumina HiSeq sequencing reads and is caused by gradual de-synchronisation of Sequencing By Synthesis (SBS) amongst the clonal molecules of a cDNA library and as such is inherent to ensemble SBS platforms. Base calls on ensemble SBS platforms are made based on the average signal from a clonal cluster of cDNA molecules (Fuller *et al.*, 2009). SBS readouts may lag behind, known as phasing, or run ahead, known as pre-phasing. Causes of phasing and pre-phasing include failure to cleave a 3-OH' blocker or fluorophore from the target cDNA and failed binding of a 3-OH' blocker respectively. Proportionately the number of cDNAs affected by phasing or pre-phasing increases with each cycle of sequencing, thereby increasing noise to signal ratio in the base-call process and thereby increasing probability of error and decreasing QPHRED (Kircher *et al*., 2011).

The majority of reads (73%) passed the quality-filtering step before being assembled into 331,569 transcripts. Sequence and assembly quality assessment metrics such as the transcript length distribution pattern, GC content, were, relatively speaking, comparable to or an improvement upon those seen in higher plants (Kalra *et al*., 2008) and other green algae, although proportion of reads remaining after trimming was lower (Ioki *et al*., 2012a; Rismani-Yazdi *et al*., 2011). The lower proportion of remaining reads may be due to more stringent quality control parameters used in this work, although parameters for other work are not known. Sequencing yield, N50, maximum transcript length, average transcript length and transcript number were superior in the *B. braunii* assembly generated in the cases where these metrics are specified for other compared studies (Guarnieri *et al*., 2011; Molnár *et al*., 2012; Sun *et al*., 2013; Yang *et al*., 2013).

The Trinity *de novo* assembly pipeline was chosen because despite being a very new program, it has been tested on Eukaryotic RNA-Seq data and outperformed the alternative assemblers, SOAPdenovo, ABYSS and Oases in nearly all assembly quality metrics compared in a recent study (Zhao *et al*., 2011). The improvement on assembly quality when Trinity was used instead of its alternatives was reflected in the increase in number of coding proteins identified. The memory requirement for Trinity was lower (57.1Gb) than the 137Gb required for Oases, which was prohibitive. Furthermore, memory requirement was correlated with data size, making memory efficiency critical for analysis of the 201 Gb of *B. braunii* sequence data in this experiment. However, the reduced memory usage was a trade-off as run-time was 20 to 100 times longer with Trinity (Zhao *et al*., 2011). Trinity was specifically created for the *de novo* assembly of short-read RNA-Seq data from NGS platforms, as opposed to

being an extension of existing genome assembly programs as in the case of alternatives. Lastly, Trinity enabled streamlined analysis as it supports RSEM (Li & Dewey, 2011), DESeq (Love *et al*., 2014) and edgeR (Robinson *et al*., 2010) - downstream differential analysis packages that account for biological and technical variation in the data.

3.4.2 Functional annotation

3.4.2.1 BLAST search of NCBI Non-Redundant Protein Database

A BLAST search of all predicted open reading frames from the assembled transcriptome enabled characterisation of 39% of the *B. braunii* assembly, leaving a considerable percentage (61%) unannotated. However, the proportion of BLAST annotations is in line with that achieved in other studies of the Chlorophyta and higher plants (Kalra *et al*., 2008; Molnár *et al*., 2012).

Transcript assignment to taxa by MEGAN revealed 202,771 unidentified transcripts with no designated taxa and 10,278 transcripts assigned to the Amoebozoa, which could represent non-botryococcal transcripts in the dataset, although may indeed partly reflect the non-axenic nature of *B. braunii* cultures and contamination introduced during sample preparation. Despite polyA-enrichment of mRNA in the library preparation, the association of a bacterial consortium with *B. braunii* cultures means that a degree of contamination by bacterial transcripts will always be present. However, the Chlorophyta are poorly represented in the NCBI BLAST database. Only nine Chlorophyta species have sequenced genomes listed on NCBI: *Helicosporidium*, an undefined *Micromonas* sp., *Micromonas pusilla*, *Volvox carteri*, *Ostreococcus lucimarinus*, *Coccomyxa subellipsoidea* C-169, *Chlorella variabilis*, *Ostreococcus tauri* and *Chlamydomonas reinhardtii*. Just 95,234 protein sequences of all 43,671,159 enlisted on the NCBI Non-Redundant database are Chlorophyta. As such it is possible that unannotated transcripts or those ascribed to other taxa are Botryococcal in origin and are rather an indicator of biases and deficiencies in the NCBI database used. Further they may be a result of localised sequence homology with phylogenetically distant organisms. The high number of proteins with no hits and unassigned proteins could also be due to the stringent BLAST parameters used and the limitation of three hits per query having an adverse effect on the LCA algorithm. Less conservative BLAST parameters would have been selected were the BLAST search not for annotation purposes rather than MEGAN analysis.

The large number of transcripts annotated by BLAST analysis under Insecta can be attributed to the presence of *B. terrestris* contamination, which was detected during

the course of differential expression analysis described in detail in subsequent chapters when unusual patterns in the data were further investigated and confirmed by sequence homology to originate from the *B. terrestris* genome. Unfortunately, this discovery was not made early enough to re-commence the BLAST search of the entire assembly. Instead all 27,930 cDNA sequences from the assembly that were reported by gmap to successfully map and align to the *B. terrestris* genome were removed from the *B. braunii* transcriptome, regardless of any score metric. The rigorous approach to removal of *B. terrestris* contamination was taken because *B. terrestris* is phylogenetically distant from *B. braunii* and therefore a high degree of confidence can be conferred in successful mapping and alignment of transcripts by gmap, which has an error rate of less than 5% (Wu & Watanabe, 2005). Furthermore, the contamination was present at a low level of 1% of all transcripts across all samples contributing to the assembly and *B. terrestris* was being worked on in the lab at the time of RNA isolation for the *B. braunii* transcriptome. These facts taken along with sequence homology, mapping and alignment data provide strong evidence for a genuine contamination event.

3.4.2.2 Gene Ontology annotation

Kinase, protein kinase and protein binding subcategories within the GOSlim2 annotation were all well represented in the *B. braunii* transcriptome. Protein kinases are conserved from bacteria to humans as components of phosphorylation induced signal transduction pathways, thus are critical to functions including cell expansion, division, apoptosis, proliferation and differentiation (Haruta *et al*., 2014; Manning, 2002; Nishida & Gotoh, 1993). The main function is phosphorylation, which leads to a change in localisation, relations with other proteins or enzyme activity. Protein kinases belong to an extensive family of proteins with catalytic domains that are highly conserved amongst types of protein kinase such as serine/threonine kinase and tyrosine too. The high numbers of GO annotations assigned to transcripts under metabolism, cell organisation and biogenesis, biosynthesis, nucleobase, signal transduction and protein modification indicate active growth, metabolism and cell communication.

Annotations relating to lipid production and lipid specific transport, which might be expected to be highly represented in an organism that produces the mass of oil that *B. braunii* is reported to, do not feature amongst the highly represented GOs. However, relating to the lipid of interest, the entire pathway is mapped and corroborates with other studies in suggesting that botryococcenes are produced via the MEP/DOXP pathway as opposed to the more commonplace route via mevalonate (Ioki *et al*., 2012b).

3.4.2.3 KEGG pathway analysis

The annotation of transcriptome-wide gene products with EC numbers, KOs and GO terms was an automated process tailored to maximise the number of confident annotations. Initially EC numbers were used to map transcripts to the KEGG metabolic pathways map but upon investigation of missing components it was decided to use KOs to ensure capture of gene products not classified as enzymes, such as protein subunits. To ensure confidence in the origins of transcripts, the annotated dataset was filtered to the *Viridiplantae* level prior to mapping to reference pathways. The absence of gene products mapped to reference pathways in which they were expected to be present in *B. braunii* may be due to the filtering process. For example, a *B. braunii* transcript mapping to the photosystem II component, psbD, was absent, despite being annotated with the corresponding KO, because it had BLAST hits with the brown alga *Saccharina japonica* and the haptophytes *Phaeocystis globosa* and *Phaeocystis antarctica*, and was binned to the lowest common ancestor.

The manual curation of annotations of pathways of interest, the basis of which was achieved through the automated process, ensured that absent components could be accounted for. The filtering process binned annotations to the lowest common ancestor where BLAST hits within 0.1% of the best hit were present in a taxon, even if there were lower scoring hits in another taxon of an equivalent level. On this basis, gene product annotations retrieved by manual curation were not necessarily disregarded.

Viridiplantae was selected as the taxonomic group to which the automated annotation was limited because higher plants diverged from one lineage of a group of three flagellated eukaryotes (Rhodopytes, Glaucophytes and *Viridiplantae*) that had diverged after the endosymbiosis of an ancestor of the cyanobacteria event (Ball, 2005; Yoon, 2004). Protein and pathway level conservation amongst the *Viridiplantae* is high (Bisova, 2005; Gutman, 2004; Hedges *et al*., 2004; Lyubetsky *et al*., 2013; Mittag, 2005). Whilst homologies with taxa outside the *Viridiplantae* may have caused the absence of pathway components that were of genuine *B. braunii* origin, the next least conservative taxonomic group to filter to would have been the Eukaryotes. Given problems encountered with fungal contamination of the dataset (discussed in section 5.2.4) the higher level of stringency was chosen to avoid inclusion of the fungi.

Where enzymes were important targets and annotation with KOs had not discovered key components, such as in the terpenoid pathway, EC numbers were used to map pathways

Many of the most abundantly mapped pathways in *B. braunii* coincided with those in *Arabidopsis thaliana*, *Oryza sativa* and *Chlorophytum borivilia,* including

Spliceosome, RNA degradation and transport, pyrimidine metabolism and protein processing in endoplasm pathways (Kalra *et al*., 2008). High abundance in these pathways indicates that *B. braunii* was at an active phase of growth during the time of harvest, which agrees with previous studies suggesting that exponential growth of *B. braunii* occurs during the initial eight days after inoculation ("Effects of harvesting method and growth stage on the flocculation of the green alga Botryococcus braunii," 1998).

Transcripts mapped to the Yeast Cell Cycle may be due to localised areas of homology in proteins caused by overlaps in cell cycle protein domains between yeast and *Viridiplantae*, as both utilise MAPK signalling and cyclin-dependent kinases (Jonak, 2002; Strickfaden *et al*., 2007).

3.4.2.4 Assignment of protein domain families using Pfam scan

Protein families have varying degrees of sequence homology, which can result in low pair-wise sequence homology. However, structure and domains are conserved. HMMs were used to assign protein families to *B. braunii* ORFs because this type of model infers conserved domains that define a family and regions that are diverse in sequence in family members.

WD40 domains are among the top 10 most abundant domains in Eukaryotic genomes and are involved in cell cycle control, transcription regulation, vesicular transport, cytoskeletal assembly and signal transduction, most commonly by ubiquitination and histone methylation, often forming a rigid platform for protein - protein interaction (Stirnimann *et al*., 2010; Xu & Min, 2011).

3.4.3 Summary

The notable presence of protein kinase and WD40 domains and gene ontologies describing vesicles and related to catalytic activity, transportation, binding and transferase activities indicates that important processes in *B. braunii* are mainly involved in metabolism and cell communication. More specifically, when taking into account the gaps in the KEGG pathway maps alongside abundant domains and activities inferred from GO terms, it becomes clear that most metabolic activity is involved in carbohydrate and energy metabolism.

The well-covered areas of the *B. braunii* KEGG pathways map were in agreement with pathway annotation in other green algae and the numbers of GO terms in the Molecular Function and Cellular Component categories were remarkably also similar. However, the number of *B. braunii* transcripts annotated with GO terms under the Biological Process category was much higher, by approximately threefold (Yang *et al*., 2013), although this may be a result of the overall higher number of GOs assigned.

The high proportion of translated ORFs from the transcriptome sequence and similarities in the representation of GO terms and KEGG pathways in the *B. braunii* transcriptome with other *Viridiplantae* species indicate that the assembly provides a comprehensive and high quality reference for differential expression analysis in subsequent chapters and for use in wider research by publication (Sun *et al*., 2013; Yang *et al*., 2013).

Chapter 3 Summary

In this chapter, 12 replicate *B. braunii* cultures were each cultured in and harvested from conditions of LD and LL over a 28 hour time series. This sampling regime was selected to ensure maximum coverage of the transcriptome by minimising any effects of circadian and diurnal changes in mRNA abundance, as documented in *Arabidopsis thaliana*. mRNA purified from harvested *B. braunii* was sequenced using Next Generation Illumina Sequencing and used to construct a diel transcriptome, *de novo*. The assembled transcriptome was comprehensively annotated and curated using TBLASTX search results, assignment of KEGG, GO and EC terms and identification of conserved domains.

CHAPTER 4

VALIDATING THE *BOTRYOCOCCUS BRAUNII* TRANSCRIPTOME

4.1 Introduction

The generation of the *Botryococcus braunii* transcriptome (Chapter 3) allows molecular and biochemical pathways to be elucidated in fine resolution. To this end, this chapter aims to use protein sequence characterisation to predict the core clock components of *B. braunii*, with the ultimate goal of generating a model of the clock oscillator and network of clock control.

The *Viridiplantae* arose approximately 1,500 million years ago (MYA) from an endosymbiotic event whereby a eukaryote entered into symbiosis with an ancestor of the cyanobacteria and the primary plastid type, directly descended from this event are those contained within the chlorophytes (rhodophytes and glaucocystophytes too) (Yoon, 2004). A second endosymbiotic event resulting in incorporation of a proteobacteria that became present- day mitochondria gave rise to all photosynthetic eukaryotes (McFadden & van Dooren, 2004). Enslavement of both the cyanobacteria and proteobacteria entailed mass gene transfer from these organisms to the host (Raymond & Blankenship, 2003). Two of the three proteins that comprise the extant cyanobacterial clock were already established within ancestral cyanobacteria lineages prior to the primary and secondary endosymbiotic events that led to the founding members of present- day *Viridiplantae*. The third is thought to have been acquired approximately 1,000 MYA (Dvornyk *et al*., 2003). If the origins of *Viridiplantae* photosynthetic machinery lie in these endosymbiotic events, it follows suit that some Chlorophyte species' circadian machinery may also, although incomplete plastid sequencing leaves this avenue so far unexplored (Ditty *et al*., 2003). So far unexplained plastid transcripts, shared metabolic pathways and prokaryote- like chromosomal traits hint at horizontal gene transfer (HGT) events but the extent of parallel HGT in the Chlorophytes is unclear, nevertheless it was little work to include the cyanobacterial core clock components to rule them out for thoroughness (Ferraz *et al*., 2006; Molnár *et al*., 2012; Raymond & Blankenship, 2003).

The pseudo-response regulators (PRRs) 1, 3, 5, 7 and 9 and the CCA1/LHY (Circadian Clock Associated 1/ Late Elongated Hypotyl) type proteins, are *Viridiplantae* core clock component domains and appear to be conserved throughout the family. Whilst other genes integral to plant clocks have been identified, these more ancestral components are well characterised lynch- pins of the circadian oscillator (Beales *et al*., 2007; Corellou *et al*., 2009; Kaldis *et al*., 2003; Murakami *et al*., 2006; Okada *et al*., 2009). In *A. thaliana*, PRR1 is known as Timing of CAB expression 1 (TOC1). In other plants species TOC1 is referred to as PRR1 therefore the same nomenclature is used in this chapter and others.

The conserved higher plant clock mechanism is currently modelled as a complex three- loop oscillator incorporating a morning loop (*i.e.* genes that are expressed around dawn), an evening loop (*i.e.* genes that are expressed around dusk) and a third connecting the former two loops. The morning loop is comprised of interactions between CCA1, LHY and the PRRs 5, 7 and 9. The evening loop is comprised of interactions between TOC1, ELF3, ELF4, GI and LUX. Finally, the third loop connects the morning and evening loops via interactions of the evening complex with PRR9 (Figure 2) (Pokhilko *et al*., 2012). In the microalga, *Ostreococcus tauri,* which has a minimised version of the higher plant clock, the CCA1/LHY type proteins are represented by CCA1 and the PRRs by just PRR1 (Figure 3) (Corellou *et al*., 2009; Thommen *et al*., 2012). *C. reinhardtii* has six plant like circadian genes including CCA1/LHY type components and some other DNA- binding proteins that are exclusive to the species, although the functional domains are similar to some higher plant clock protein domains (Matsuo & Ishiura, 2011). However, a model of the *C. reinhardtii* circadian elements as a functioning clock has not been established. The Kai proteins, A, B and C are the proteins underpinning cyanobacterial circadian oscillators (Kondo, 2007).

In this chapter, a multi-pronged approach to identification of clock components in the *B. braunii* assembly was taken. Using the *B. braunii* ORF database generated in Chapter 3, *B. braunii* predicted clock components were initially identified by protein sequence homology with known clock components of *Arabidopsis thaliana*, *Ostreococcus taurii*, *Chlamydomonas reinhardtii* and *Synechococcus elongatus*.

The focus of the investigation was directed towards the core clock proteins; the pseudo-response regulators and LHY/CCA1 type components from *Viridiplantae*, and domain conservation was used to confirm protein sequence homology results using both sequence and Hidden Markov Model- based search terms. Multiple sequence HMM alignments were generated between the best *B. braunii* candidates and their respective model clock counterparts. Using cross- referenced data from studies of *A.*

thaliana, *O. taurii*, *C. reinhardtii* (Murakami *et al*., 2006; Pokhilko *et al*., 2012; Troein *et al*., 2011) the complex oscillator of higher plants (Figure 2) and the simpler mechanism of green algae (Figure 3) are used as references in order to propose a clock model for *B. braunii*.

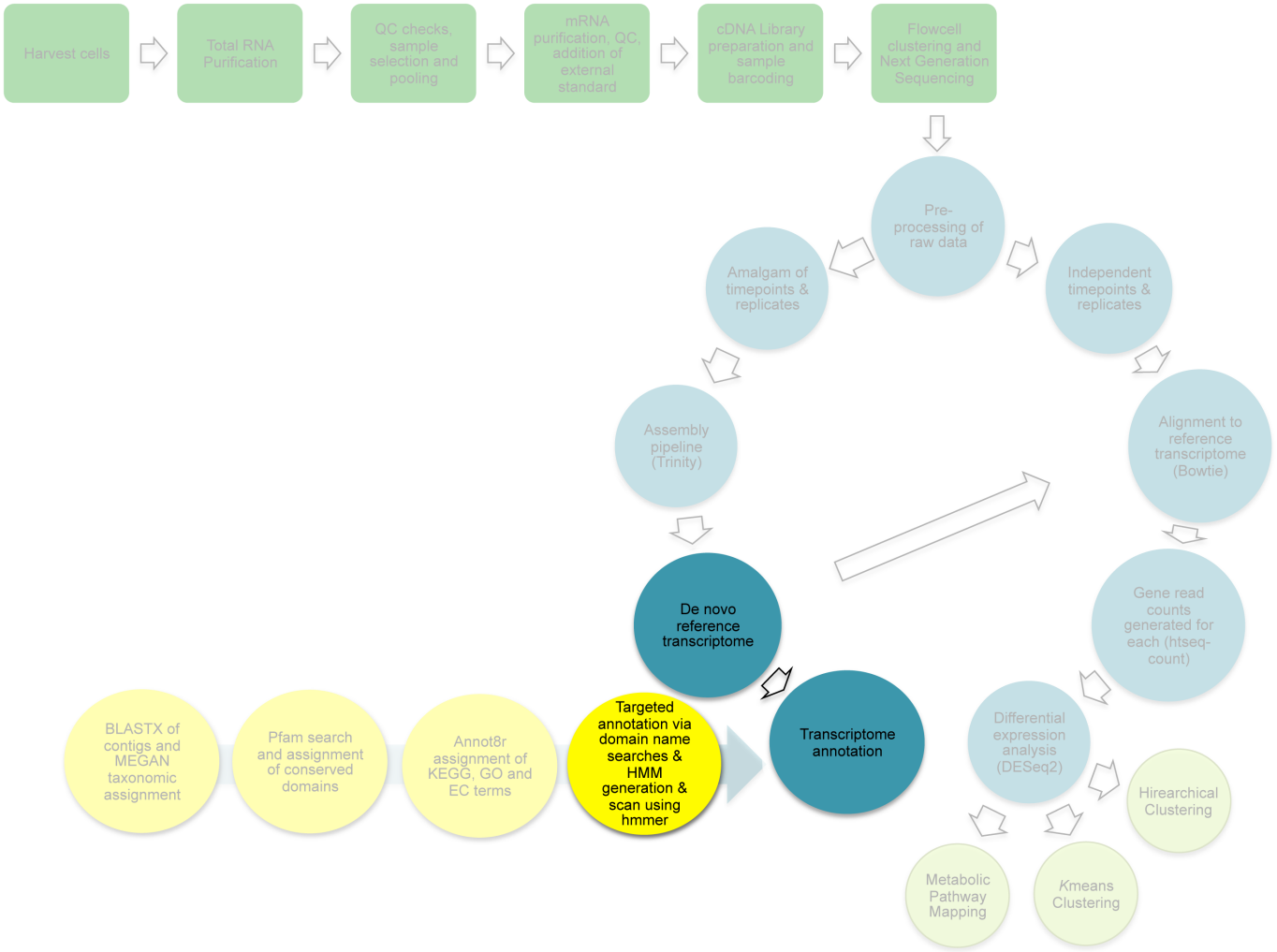


Figure 1 Workflow of experimental and bioinformatic methods

Practical laboratory methods (green) through to bioinformatic analysis (blue) used to investigate the transcriptomics of *B. braunii* over a 28-hour time series under two different photo-regimes. This chapter describes the identification and characterisation of predicted *Botryococcus braunii* clock components (yellow) sequence. Faded out components are addressed in Chapters 3 and 5.

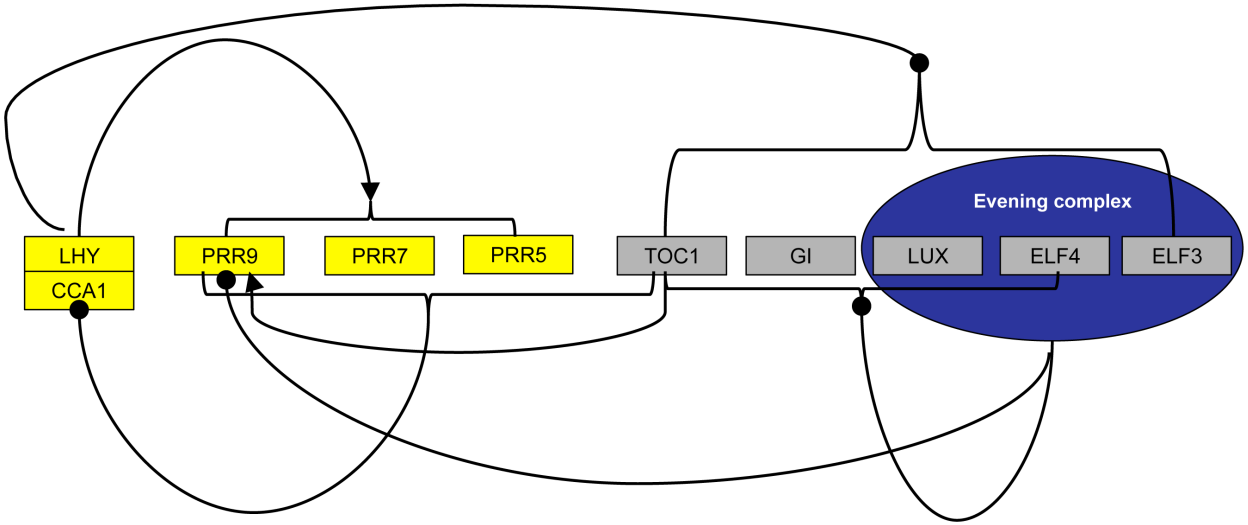


Figure 2 Higher plant clock model

A model of the higher plant clock based on that proposed by Pokhilko *et al* 2012. Proteins only are shown for simplicity. Morning phased elements are shown in yellow boxes and evening phased elements in grey boxes. The association of LUX, ELF4 and ELF3 into an 'evening complex' is denoted by a blue oval. Transcriptional regulation is shown by solid black lines, terminating in a solid circle representing down- regulation or a solid arrowhead for up regulation.

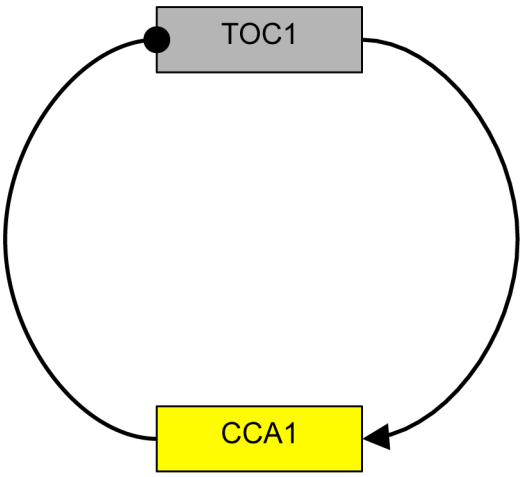


Figure 3 *Ostreococcus tauri* clock model
The proteins of the simple clock mechanism of *O. tauri* are shown with the morning component CCA1 in yellow and the evening component TOC1 in grey. TranscriptionI regulation is shown by solid black lines terminating with a solid circle for down- regulation and a solid arrowhead for up- regulation.

4.2 Materials and Methods

4.2.1 Identification of clock protein homologs by sequence

The literature was mined for known clock component protein sequences from the *Viridiplantae* species; *A. thaliana, O. tauri, C. reinhardtii* and the cyanobacterium; *S. elongatus (*Table 1*)*. The protein sequences of the clock components from model organisms were downloaded and saved as FASTA format files.

A BLAST protein database was generated from a FASTA file of the *B. braunii* transcriptome ORFs (Chapter 3) using default parameters of makeblastdb – a command line tool available as part of the BLAST+ toolkit (Camacho *et al*., 2009). The FASTA files containing known clock protein sequences from model organisms were used as queries in a local, standalone multi- thread (8 CPUs were used) search of the *B. braunii* transcriptome. The results of the BLAST search were parsed to produce a tabulated list of top hits for each clock protein using the custom Perl script, blast_parse_pw2tab.pl (courtesy of University of Exeter Microbial Biofuels Bioinformatics group).

For each clock protein query, the protein sequence of the top hit *B. braunii* ORF was used in a reciprocal BLAST search against the NCBI non-redundant protein database (Pruitt *et al*., 2013) using the NCBI online search facility (http://blast.ncbi.nlm.nih.gov/). *B. braunii* ORFs with reciprocal BLAST hits from non-*Viridiplantae* were discarded and the next best *B. braunii* ORF hit against the model clock protein (e value < $1e^{-5}$) was used in the reciprocal BLAST search.

Organism and query protein	Accession
Arabidopsis thaliana	
Timing of CAB expression 1 (TOC1)	AT5G61380 (TAIR)
Pseudo Response Regulator 3 (PRR3)	AT5G60100 (TAIR)
Pseudo Response Regulator 5 (PRR5)	AT5G24470 (TAIR)
Pseudo Response Regulator 7 (PRR7)	AT5G02810 (TAIR)
Pseudo Response Regulator 9 (PRR9)	AT2G46790 (TAIR)
Late Elongated Hypotyl (LHY)	AT1G01060 (TAIR)
Circadian Clock Associated 1 (CCA1)	AT2G46830 (TAIR)
LUX	AT3G46640 (TAIR)
Ostreococcus tauri	
TOC1	AY740079 (Genbank)
CCA1	AAU14271.1 (Genbank)
Chlamydomonas reinhardti	
CHLAMY1	GI:159476646 (RefSeq)
Synechococcus elongates	
KaiA	GI:22002533 (DDBJ)
KaiB	GI:22002534 (GenBank)
KaiC	GI:22002535 (GenBank)

Table 1 Clock protein BLAST queries

Model clock protein sequences used to BLAST search against the *Botryococcus braunii* ORF database, with accession codes and databases in which they are deposited in brackets.

4.2.2 Identification of clock protein homologs by functional domain

4.2.2.1 Domain and motif identification

The conserved domains of model clock components (Table 1) were identified using the Pfam A protein family database online search facility (http://pfam.xfam.org/) (Finn *et al*., 2009). Also using the Pfam online facility, protein sequences of *B. braunii* ORFs that had sequence homology with known clock proteins and had resulted in a BLAST hit from an organism within the *Viridiplantae* (section 4.2.1 Identification of clock protein homologs by sequence) were searched against the Pfam A database to identify conserved domains.

4.2.2.2 MUSCLE alignment of Botryococcus braunii *putative clock components to known clock components*

Using MUSCLE (Edgar, 2004), multiple protein sequence alignments of predicted *B. braunii* ORFs, identified by sequence homology and conserved domain architecture with model clock proteins were generated from FASTA files containing protein sequence data. Alignments were visualised with the Seaview package, Version 4.5.2 (Galtier *et al*., 1996).

4.2.2.3 HMM alignment of Botryococcus braunii *predicted clock components to model clock components*

HMMs of the conserved domains of model clock proteins were downloaded from Pfam (http://pfam.xfam.org/) and the command line tool hmmalign from the HMMER (V.3) toolkit was used with default parameters to align the protein sequences of predicted *B. braunii* and model clock components to the downloaded HMMs.

4.2.2.4 Keyword search of HMM Botryococcus braunii *database*

Conserved domains of core clock proteins in model organisms were used as search terms with the command line tool, grep, to identify *B. braunii* ORFs containing the same domains in the *B. braunii* HMM profile database generated from the *B. braunii* transcriptome (section 3.2.4.3).

4.2.2.5 Clock component HMM generation and transcriptome HMM scan

Using the default parameters of the command line tool, hmmbuild from the HMMER (V.3) toolkit, custom HMMs were generated for core clock proteins from MUSCLE protein alignments. The Muscle protein alignments to proteins of model organisms were generated in the same way as described in section 4.2.2.2, excluding *B. braunii* transcripts from the input FASTA file.

Hidden Markov Models create profiles of conserved domains within proteins, emitting position specific probabilistic inferences of amino acids based upon a set of model “training sequences”. HMMs are constituted of match states, insert states, delete states and transmission states. Match states are representative of points of conservation in the alignment. In match states, “probable” AA residues are emitted by the model based upon observed counts of the residues in a column of an alignment and the pseudo- counts given by a prior. Insert states allow for the presence of residues between match states, emission probability of insert states is calculated from Pfam residue frequency data. Delete states are placeholders where a residue is not emitted. Transmission states give the probability of moving from one of the aforementioned states to another and are calculated based upon normalised observed counts of the alignment and a prior (Johnson, 2006).

B. braunii predicted ORFs containing conserved clock protein domains were searched for within the transcriptome, using hmmscan (also from HMMER V.3) on the command line and the generated HMM files of known clock proteins as queries.

4.3 Results

4.3.1 Identification of clock protein homologs by sequence

Core clock components from *A. thaliana*, *O. tauri* and *C. reinhardtii* all resulted in the discovery of *B. braunii* transcripts with sequence homology above the defined cutoff (BLAST E value ≤ $1e^{-05}$) (Table 2). Clock sequence homology results for the pseudo-response regulator family; *A. thaliana* TOC1 and PRRs 3, 5, 7 and 9 all lead to only *B. braunii* transcript, comp56012_c0_seq2 (E values between $1.00e^{-24}$ and $6.00e^{-33}$), which was from this point onwards considered a predicted pseudo response regulator and used to represent the family in subsequent analysis. For simplicity, comp56012_c0_seq2 was named *BbPRR* from this point onwards.

LHY and CCA1 from *A. thaliana* shared sequence homology with transcripts comp170985_c0_seq11 (E value = $1.00e^{-23}$) and comp149716_c0_seq1 (E value = $3e^{-22}$) respectively. However the reciprocal BLAST result for transcript comp149716_c0_seq1 was an Myb DNA binding domain containing protein from the cellular slime mould *Polysphondylium pallidum* of the Amoebozoa Kingdom (E value = $7e^{-74}$). As *P. pallidum* is not from within the *Viridiplantae*, transcript comp149716_c0_seq1 was discounted and replaced with the next best BLAST hit. The next best hit was comp170985_c0_seq11 ($1.00e^{-21}$); the same transcript as the best hit for *A. thaliana* and *O. tauri* CCA1. From this point onwards comp170985_c0_seq11 was named *BbCCA1.*

comp56565_c0_seq2 shared sequence homology with *A. thaliana* LUX (E value = $4.00e^{-22}$). Homologs for the *O. tauri* pseudo response regulator TOC1 and CCA1 were found in *B. braunii* transcripts comp112899_c0_seq2 (E value = $5.00e^{-68}$), and BbCCA1 (E value = $8.00e^{-24}$). CHLAMY1, the *Chlamydomonas reinhardtii* clock component showed sequence similarity with comp166095_c2_seq1 (E value = $2.00e^{-73}$).

The sequences of comp65295_c0_seq1 and comp179979_c0_seq1 demonstrated homology with those of ELF3 and ELF4 with E values of $2.00e^{-65}$ and $5e^{-13}$ respectively.

Transcripts BbPRR, BbCCA1, comp56565_c0_seq2, comp112899_c0_seq2 and comp166095_c2_seq1 had reciprocal BLAST search results with E values above the selected cut- off of e^{-05} from organisms within the green algae, namely *Volvox carteri, Micromonas* sp. RCC299, *Coccomyxa subellipsoidea* and were thus retained for subsequent analyses. *B. braunii* comp65295_c0_seq1 and comp179979_c0_seq1 also had reciprocal BLAST search results from with another green algal species, *Coccomyxa subellipsoidia* (E values of $3e^{-75}$ and $1.00e^{-18}$ respectively). However, these

proteins were not included in further analysis because their sole purpose of inclusion in the BLAST search was to guide focus toward a likely core clock mechanism.

Cyanobacterial clock components from *S. elongatus* either did not yield sequence homologies in the *B. braunii* transcriptome above the E value cut- off or where there was a match, the reciprocal BLAST resulted in hits from outside the *Viridiplantae* clade. KaiA and KaiC protein sequences resulted in low sequence homology with comp172797_c5_seq1 (E value = 1.80) and comp333351_c0_seq1 (E value = 0.05) respectively. KaiB and transcript comp49501_c0_seq1 showed good sequence homology ($3.00e^{-026}$) but the reciprocal BLAST of comp49501_c0_seq1 resulted in a match with *Flavobacterium johnsoniae* UW101 (E value = $1e^{-46}$). Cyanobacterial clock components were thus discounted from further analyses.

Query	Hit transcript & E number	Reciprocal blast hit & E number
Arabidopsis thaliana		
TOC1	BbPRR, $1e^{-24}$	hypothetical protein VOLCADRAFT_69846 [*Volvox carteri f. nagariensis*], $2e^{-38}$
PRR3	BbPRR, $1.00e^{-027}$	See above
PRR5	BbPRR, $2.00e^{-028}$	See above
PRR7	BbPRR, $6.00e^{-033}$	See above
PRR9	BbPRR, $9.00e^{-029}$	See above
LHY	BbCCA1_c0_seq11, $1e^{-23}$	predicted protein [*Micromonas sp.* RCC299], $2e^{-29}$
CCA1	comp149716_c0_seq1, $3e^{-22}$	myb domain-containing protein [*Polysphondylium pallidum* PN500], $7e^{-74}$
	BbCCA1_c0_seq11, $1.00e^{-021}$	predicted protein [Micromonas sp. RCC299], $2e^{-29}$
LUX	comp56565_c0_seq2, $4e^{-22}$	CheY-like protein [*Coccomyxa subellipsoidea* C-169], $1e^{-118}$
ELF3	comp65295_c0_seq1, $2.00e^{-65}$	WD40 repeat-like protein, partial [Coccomyxa subellipsoidea C-169], $3e^{-75}$
ELF4	comp179979_c0_seq1, $5e^{-13}$	DUF1313 containing protein, coccomyxa subellipsoidia C-169, $1.00e^{-18}$
Ostreococcus tauri		
TOC1	comp112899_c0_seq2, $5.00e^{-68}$	WD40 repeat-like protein, partial [*Coccomyxa subellipsoidea* C-169], $9e^{-119}$
CCA1	BbCCA1_c0_seq11, $8.00e^{-024}$	As above
Chlamydomonas reinhardti CHLAMY1	comp166095_c2_seq1, $2.00e^{-073}$	Hypothetical protein COCSUDRAFT_64249 [*Coccomyxa subellipsoidea* C-169], $2e^{-120}$
Synechococcus elongatus		
KaiA	comp172797_c5_seq1, 1.8	N/A, E value > e^{-05}
KaiB	comp49501_c0_seq1, $3.00e^{-026}$	KaiB domain-containing protein [*Flavobacterium johnsoniae* UW101], $1e^{-46}$
KaiC	comp333351_c0_seq1, 0.048	N/A, E value > e^{-05}

Table 2 BLAST protein sequence homology results

Known clock components from model organisms are shown in column one with the corresponding top hits and BLAST E values in *B. braunii* column two. Results from the BLAST search of *B. braunii* clock protein homologs against the NCBI non- redundant database are shown in column three. Results with E values above the cutoff of $1e^{-05}$ are grayed out.

4.3.2 Identification of clock protein homologs by functional domain

4.3.2.1 Domain and motif identification

The conserved domains of PRRs 1 (TOC1), 3, 5, 7 and 9, LHY and CCA1 from *A. thaliana* and *O. sativa*, LUX from *A. thaliana*, PRR1 and CCA1 from *O. tauri* and CHLAMY1 subunit C1 from *C. reinhardtii* (Table 1 for accessions) were identified using the Pfam A protein family database online search facility (http://pfam.xfam.org/) (Finn *et al*., 2009). *O. sativa* homologs to TOC1 (PRR1), the PRRs 3, 5, 7 and 9 and the LHY/CCA1- type proteins were added to the original model clock components list, to compare inter- species variation in sequence length and domain position and size.

B. braunii transcript BbPRR was identified by Pfam using an E value cutoff of 1, to share response regulator receiver (RRR) and CCT domains with PRRs 1 (and TOC1), 3, 5, 7 and 9 from *A. thaliana* and *O. sativa*, and PRR1 from *O. tauri*. *B. braunii* BbCCA1 shared the Myb domain with *A. thaliana* LHY and CCA1, LHY from *O. sativa* and CCA1 from *O. tauri* (Table 3).

Clock component query organism, protein & domains	*Botryococcus braunii* domains
Arabidopsis thaliana	
TOC1; RRR domain, CCT motif	BbPRR; RRR domain, CCT motif
LHY; Myb_DNA_Binding	BbCCA1_c0_seq11; Myb_DNA_Binding
CCA1; Myb_DNA_Binding	See above for BbCCA1
LUX Myb_DNA_Binding	comp56565_c0_seq2; RRR domain, Myb_DNA_Binding
PRR3; RRR domain, CCT motif	See above for BbPRR
PRR5; RRR domain, CCT motif	See above for BbPRR
PRR7; RRR domain, CCT motif	See above for BbPRR
PRR9; RRR domain, CCT motif	See above for BbPRR
Oryza sativa	
TOC1 RRR domain, CCT motif	See above for BbPRR
LHY Myb_DNA_Binding	See above for BbCCA1_c0_seq11
CCA1 Myb_DNA_Binding	See above for BbCCA1_c0_seq11
Ostreococcus tauri	
TOC1 RRR domain, CCT motif	Comp112899_c0_seq2; Acyl CoA binding domain, Kelch_1, Kelch_4, Kelch_3
CCA1 Myb_DNA_Binding	BbCCA1_c0_seq11; Myb_DNA_Binding
Chlamydomonas reinhardti	
CHLAMY1 3 KH repeats, WW domain	comp166095_c2_seq1; 3 KH repeats, WW domain

Table 3 Model clock component and *B. braunii* sequence homolog domains
In the first column are shown the model clock components with which *B. braunii* ORFs (column two) showed sequence homology and their domains. Domains within the *B. braunii* sequence homologs are shown in column two.

4.3.2.2 Domain architecture comparison with model clock components

Model clock components that shared domains with *B. braunii* sequence homologs (Table 3) were compared with their *B. braunii* counterparts (BbPRR, BbCCA1, comp56565_c0_seq2 and comp166095_c2_seq1) by the generation of schematic diagrams mapping the alignment coordinates of Pfam domain searches (4.2.2.1 Domain and motif identification) using DOG software (Ren *et al*., 2009).

BbPRR shared both the RRR domains and CCT motifs conserved in PRRs 1 (including TOC1), 3, 5, 7 and 9 (Figure 4(a), Figure 5 (a), Figure 6 (a), Figure 7 (a), Figure 8 (a). However *B. braunii* BbPRR was 759 amino acids (AA) in length- the longest of *A. thaliana* TOC1 (618 AA), *O. sativa* PRR1 (526) and *O. tauri* PRR1 (588). The RRR domain alignment length was 111 AA in *B. braunii* BbPRR, in TOC1 and all PRRs of the three model organisms, with exception of PRR5 in *O. sativa*, which was just 72 AA. CCT domain length (43 AA) of all PRRs and TOC1 remained consistent between model organisms and in *B. braunii* BbPRR. *B. braunii* BbPRR also showed the same relative domain positioning as the model clock components, with the RRR domain near the N- terminus and the CCT domain near the C- terminus. The RRR domains were situated starting between 39 and 161 AA from the N- terminus, except for PRR5 from *O. sativa*, which started just 9 AA from the N- terminus. *B. braunii* BbPRR CCT domain began 677 AA from the N- terminus and ended 39 AA from the C- terminus.

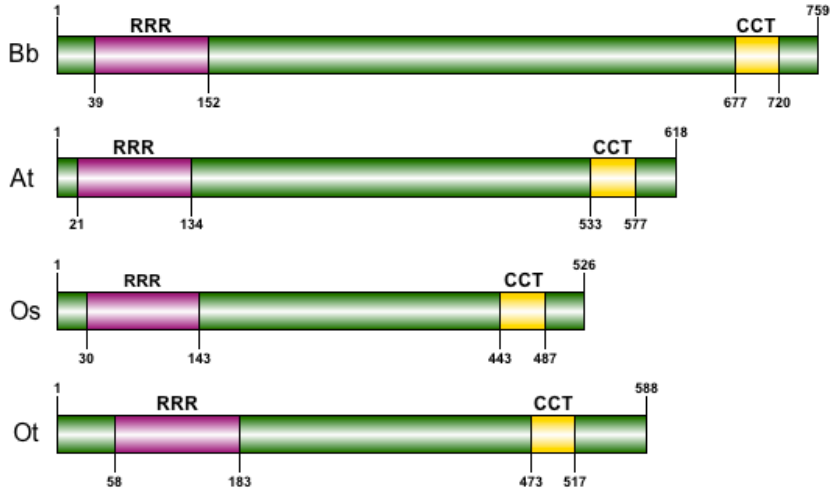
The sequence of BbPRR was aligned to HMMs of the RRR and CCT domains of PRRs 1 (TOC1), 3, 5, 7 and 9 from *A. thaliana* and *O. sativa*, and PRR1 from *O. tauri*, all of which it shared RRR and CCT domains with. *B. braunii* BbCCA1 was aligned to HMMs of the Myb DNA binding domain of LHY and CCA1 of *A .thaliana*, LHY from *O. sativa* and CCA1 from *O. tauri*.

The accuracy of the alignment of sequences to the generated HMMs was assessed by the calculation of a posterior probability score for each AA in each sequence and generation of a consensus score for each column of the alignment. The posterior probability is summarised as $P(x_i \sim y_j\ a^* | x, y)$, where xi and yj are particular positions within the sequences x and y and a* is the alignment of x and y to the HMM (Do, Mahabhashyam, Brudno, & Batzoglou, 2005). The posterior probability scores of 0- 9* equate to 0 = 0.00- 0.05, 1 = 0.05- 0.15, 2 = 0.15- 0.25, 3 = 0.25- 0.35, 4 = 0.35- 0.45, 5 = 0.45- 0.55, 6 = 0.55- 0.65, 7 = 0.65- 0.75, 8 = 0.75- 0.85, 9 = 0.85- 0.95, * = 0.95- 1.00.

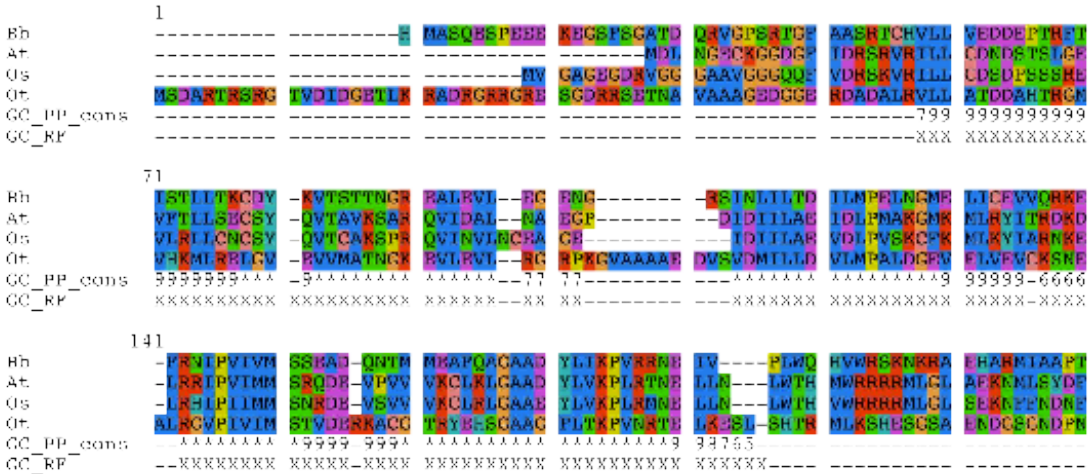
The amino acid sequences of the RRR and CCT domains in BbPRR, TOC1 from *A. thaliana* and PRR1s from *O. sativa* and *O. tauri* were highly conserved with posterior

probability scores of accurate alignment of 9 or * in the majority of positions (Figure 4 (b) and (c)). *O. sativa* PRR1 insertions of asparigine and cysteine in positions 97 and 98 preceded a short regions of four amino acids with lower conservation and lower posterior probability scores of 7 in the RRR domain (Figure 4 (b)). Following the 4 residue long region of lower alignment score, *O. tauri* PRR1 exhibited an insert unshared by *B. braunii*, *A. thaliana* and *O. sativa* in the RRR domain between alignment positions 103 and 113. Residue conservation tailed off with posterior probability score decreasing from 9 to 6 in the last four positions of the RRR domain alignment and from 9 to 5 in the CCT domain alignment (Figure 4 (b) and (c)).

(a)



(b)



(c)

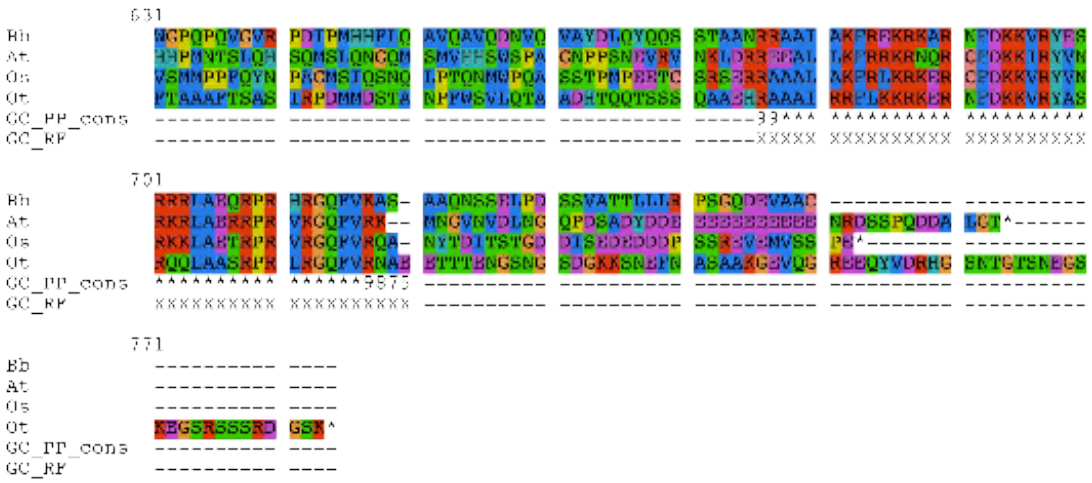


Figure 4 PRR1/ TOC1 domain architecture and alignments

(a) Schematic representation of *BbPRR* (Bb) domain structure compared with that of *A. thaliana* TOC1 (At), *O. sativa* PRR1 (Os) and *O. tauri* PRR1 (Ot) and corresponding HMM alignments of the response regulator receiver (RRR) (b) and CCT (c) domains. On the 4th row is the posterior probability score of 0- 9*, where *= 0.95- 1.00. On the 5th row the HMM location is denoted by X.

BbPRR alignment with *A. thaliana* and *O. sativa* PRR3s to the HMM of PRR3 showed the sequences were highly conserved throughout the RRR (Figure 5 (b)) and CCT (Figure 5 (c)) domains. The great majority of positions in the alignment were assigned the maximum consensus posterior probability score of * in both domains. Aside from lower posterior probability scores at the start and end of the HMM, there were short regions of lower conservation and lower posterior probability scores at positions 90 to 93, 105 to 108 and 135 to 140 (inclusive) in the RRR domain alignment. All regions with lower posterior probability score were located proximal to the site of a gap or insertion such as those of *B. braunii* arginine and valine insertions in AA positions 110 and 134. The conservation of AA residues tailed off with posterior probability score decreasing from 9 to 6 in the last four positions of the RRR and the CCT domain alignment. In position 139 of the RRR domain alignment threonine and isoleucine residues are present in the *A. thaliana* and *O. sativa* sequences and are insertions relative to the sequence of *B. braunii*.

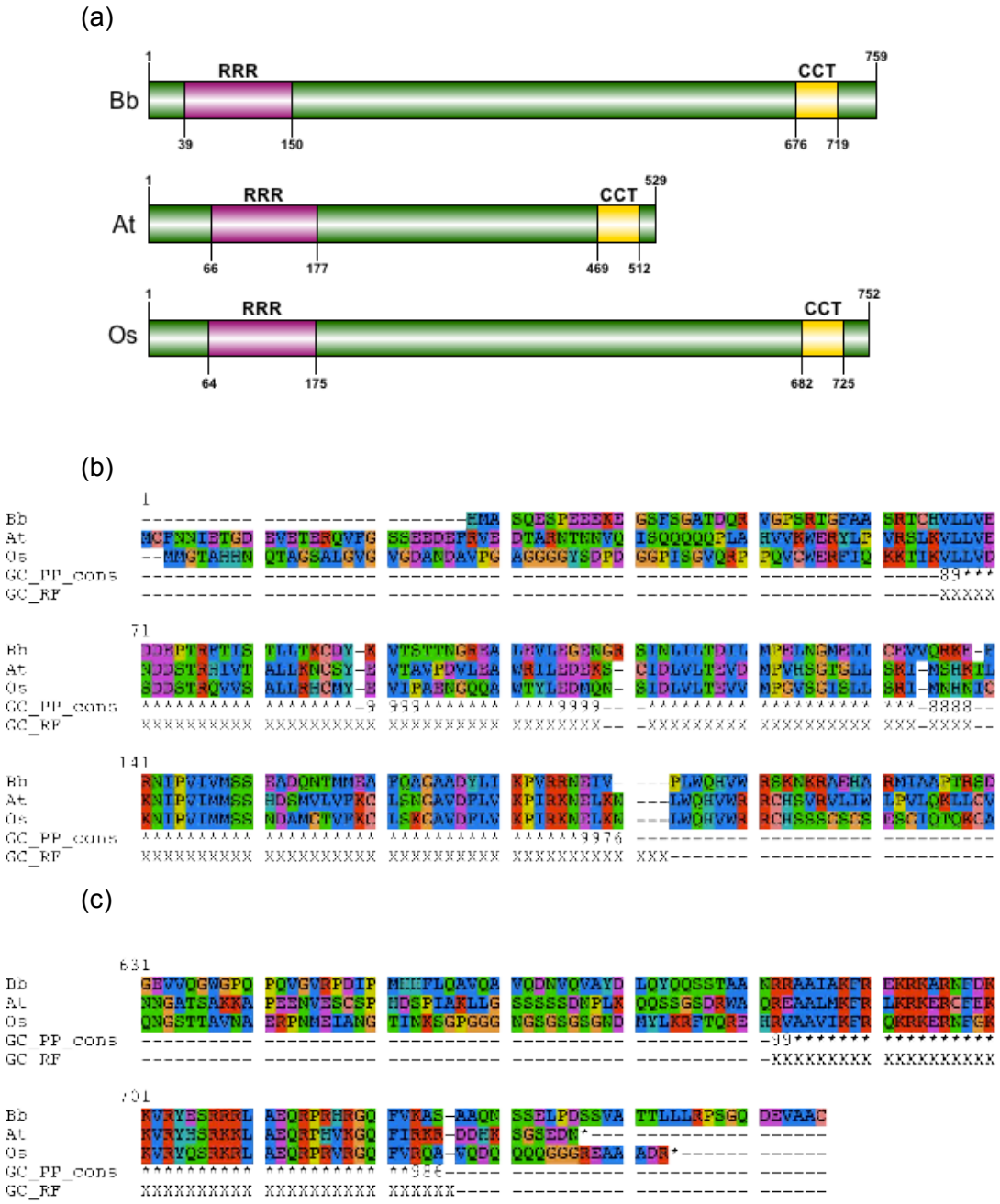


Figure 5 PRR3 domain architecture and alignments

(a) Schematic representation of BbPRR (Bb) domain structure compared with that of *A. thaliana* (At), *O. sativa* (Os), and *O. tauri* (Ot) PRR3s and corresponding HMM alignments of the response regulator receiver (RRR) (b) and CCT (c) domains. On the 4th row is the posterior probability score of 0- 9*, where *= 0.95- 1.00. On the 5th row the HMM location is denoted by X.

The AA sequences of BbPRR and *A. thaliana* and *O. sativa* PRR5s were highly conserved within the RRR and CCT domains although in the RRR domain short regions of lower conservation were associated with insertion sites (Figure 6 (b) and (c)). In positions 100- 103 of the RRR domain alignment one region of lower conservation with a lower posterior probability score of 6- 8, preceded an insertion site of an arginine residue in *B. braunii*. Another short section of lower scoring alignment (posterior probability scores of 7) was preceded by insertion of a valine residue in the BbPRR sequence at position 229 of the alignment of the RRR domain. Further along the alignment than the valine insertion, toward the C terminus is a single insertion site of an isoleucine and a glutamic acid in the sequences of *A. thaliana* and *O. sativa* respectively. *B. braunii* and *A. thaliana* contain an insertion site 40 AA in length in the RRR domain, that is not present in the *O. sativa* PRR5 sequence. Furthermore the 40 AA insertion site is not present in any other PRR protein aligned here.

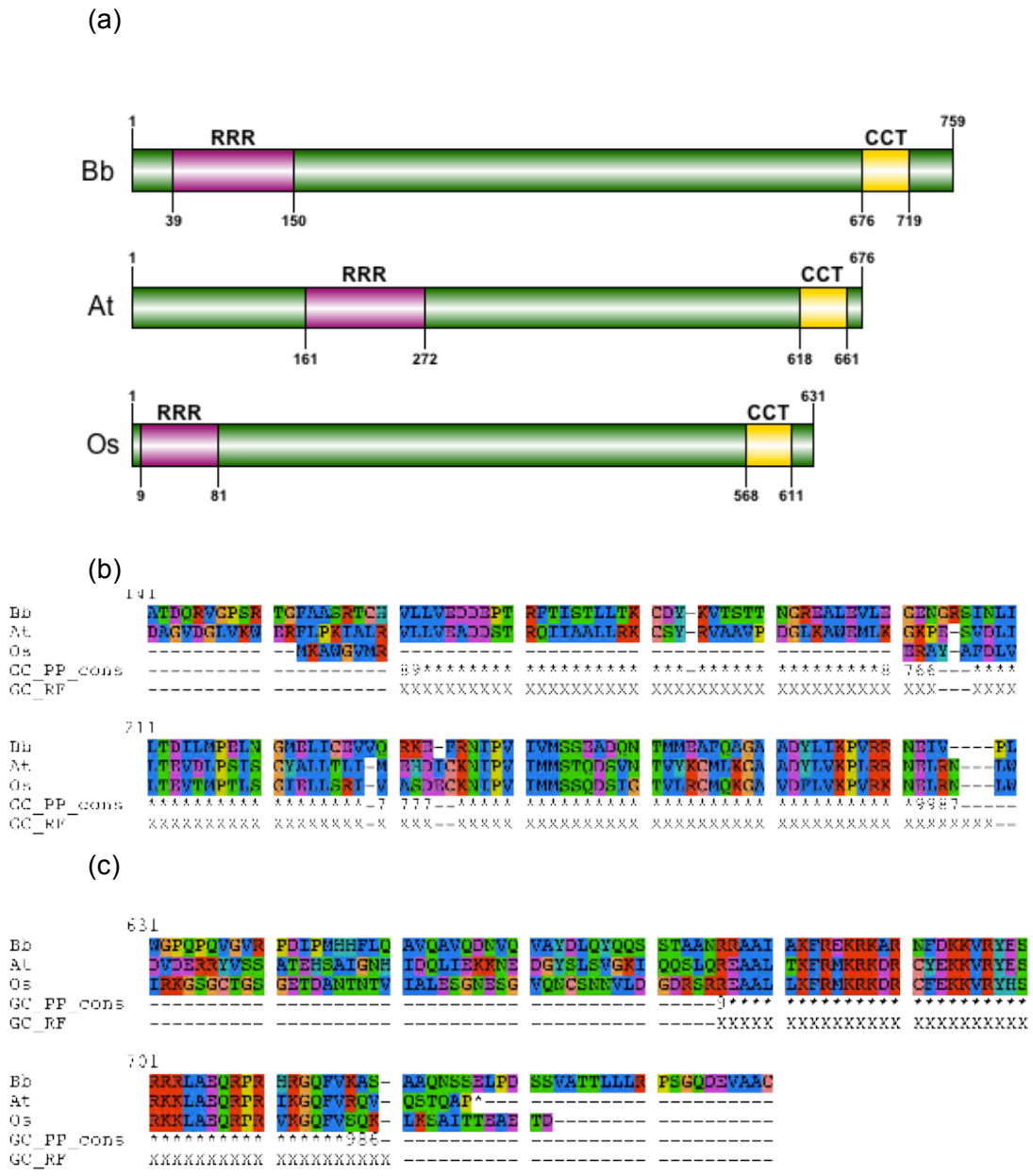


Figure 6 PRR5 domain architecture and alignments

(a) Schematic representation of BbPRR (Bb) domain structure compared with that of *A. thaliana* (At), *O. sativa* (Os), and *O. tauri* (Ot) PRR5s and corresponding HMM alignments of the response regulator receiver (RRR) (b) and CCT (c) domains. On the 4th row is the posterior probability score of 0- 9*, where *= 0.95- 1.00. On the 5th row the HMM location is denoted by X.

PRR7 protein sequences from *A. thaliana* and *O. sativa* and BbPRR were highly conserved in the RRR and CCT domains. Similar to comparison with other PRR proteins, insertion of an arginine and a valine residue in the BbPRR sequence is associated with areas of more variable sequence conservation and lower posterior probability score of alignment (7- 8) than that seen for the majority of positions (*). In the second line of the alignment or the RRR domain, *A. thaliana* and *O. sativa* contain single amino acid insertions of serine and isoleucine respectively and relative to BbPRR. The CCT domain is consistently highly conserved throughout *B. braunii*, *A. thaliana* and *O. sativa* with a high posterior probability score of * between the beginning and end clusters of lower score (between 9 and 6), consistent with HMM alignments of the other PRR proteins.

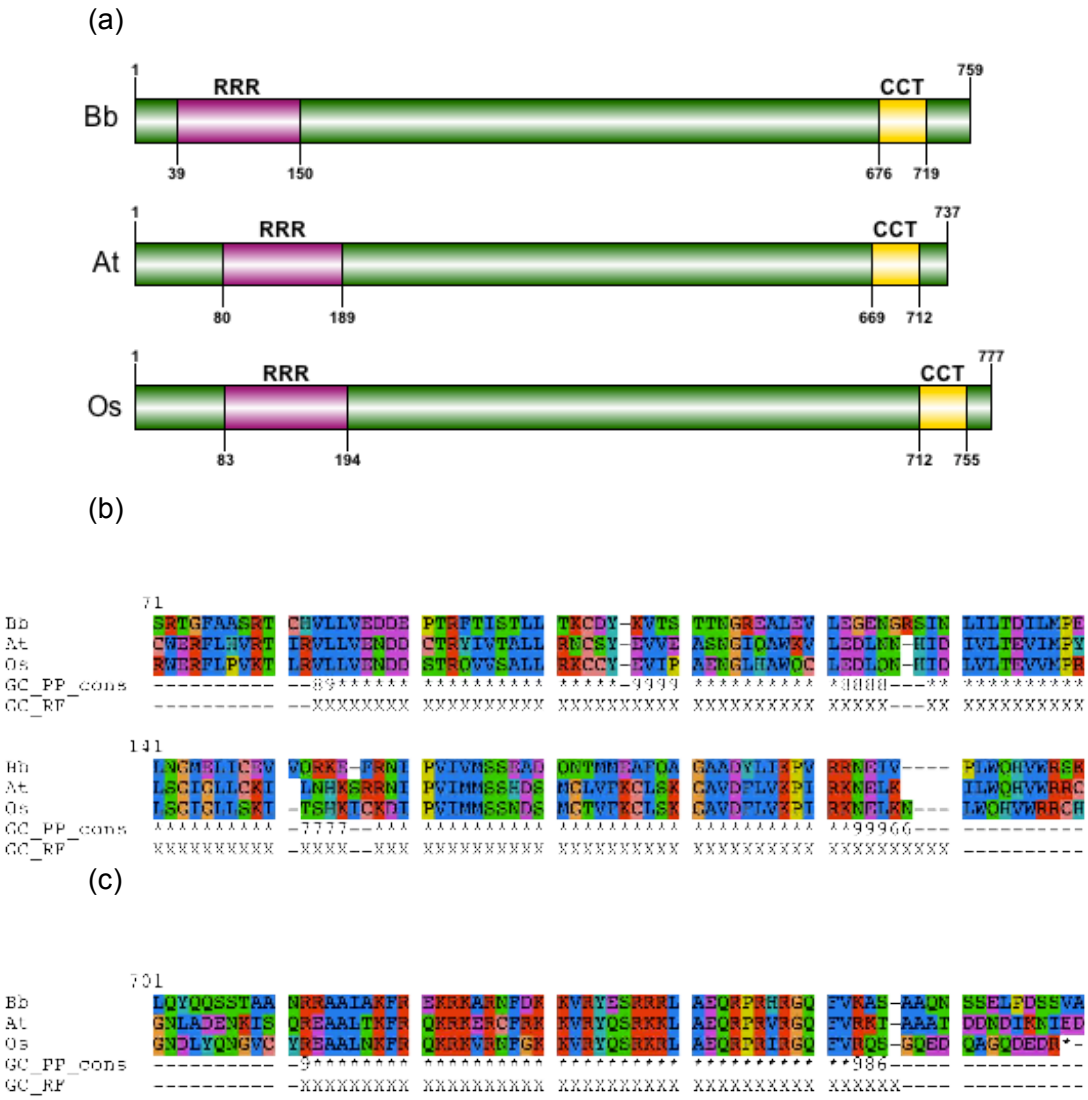
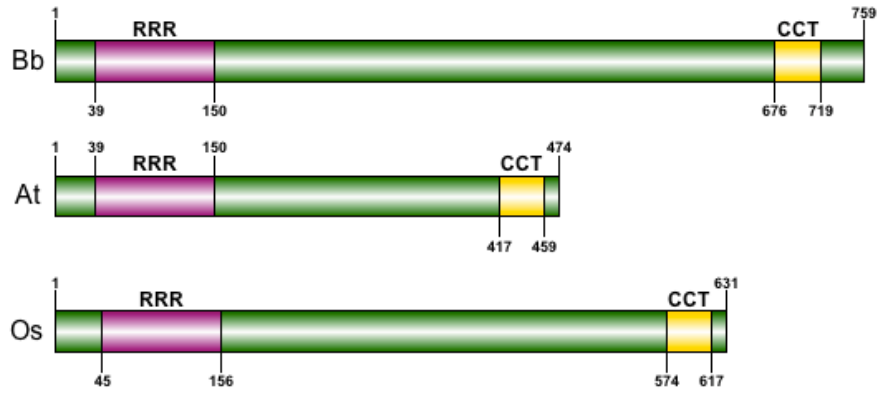


Figure 7 PRR7 domain architecture and alignments

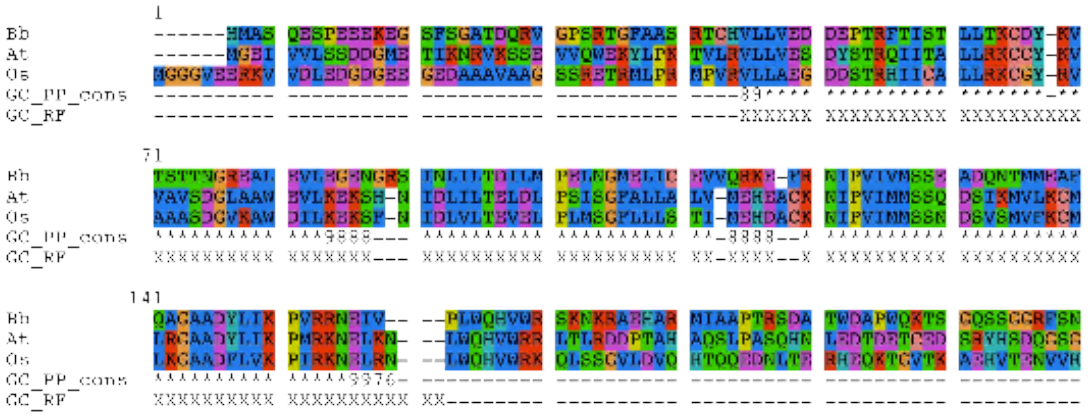
(a) Schematic representation of BbPRR (Bb) domain structure compared with that of *A. thaliana* (At), *O. sativa* (Os), and *O. tauri* (Ot) PRR7s and corresponding HMM alignments of the response regulator receiver (RRR) (b) and CCT (c) domains. On the 4th row is the posterior probability score of 0- 9*, where *= 0.95- 1.00. On the 5th row the HMM location is denoted by X.

RRR and CCT domain sequences are highly conserved in the PRR9 proteins from *A. thaliana* and *O. sativa* and also when compared BbPRR. In the RRR domain, lower alignment score (8- 9) and variability in sequence conservation is associated with insertions of arginine at position 89 and valine residues at position 113 in the BbPRR sequence that are not present in those of *A. thaliana* nor *O. sativa*. However, the RRR domains of *A. thaliana* and *O. sativa* PRR5 have an additional alanine in position 118 that is not present in BbPRR.

(a)



(b)



(c)

Figure 8 PRR9 domain architecture and alignments

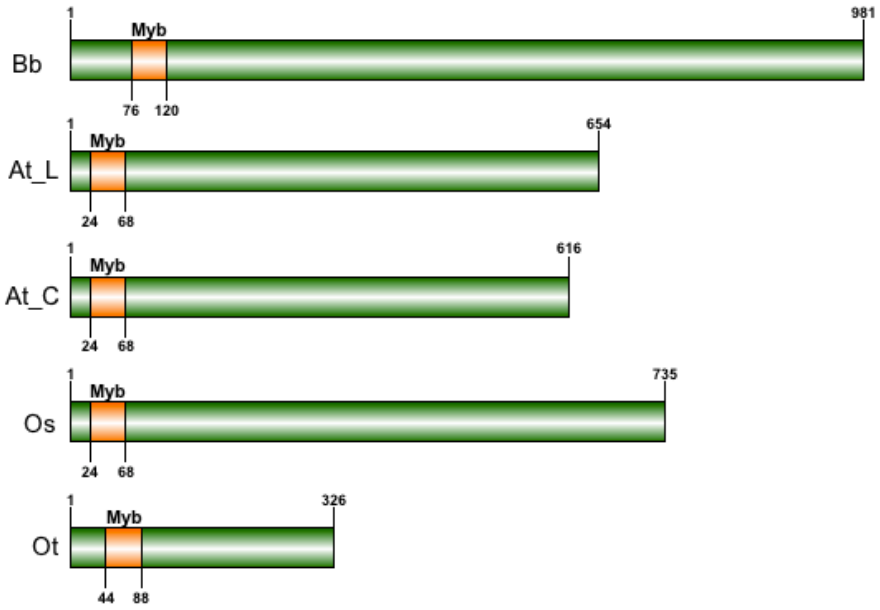
(a) Schematic representation of *B. braunii* comp56012_c0_seq2 (Bb) domain structure compared with that of *A. thaliana* (At), *O. sativa* (Os), and *O. tauri* (Ot) PRR5s and corresponding HMM alignments of the response regulator receiver (RRR) (b) and CCT (c) domains. On the 4th row is the posterior probability score of 0- 9*, where *= 0.95-1.00. On the 5th row the HMM location is denoted by X.

BbCCA1 was 981 AA- the longest of *A. thaliana* LHY (654 AA) and CCA1 (616 AA), *O. sativa* LHY (735 AA) and *O. tauri* CCA1 the shortest protein of just 326 AA. In BbCCA1 and the model clock proteins there was only one conserved domain detected by Pfam using an E value cutoff of 1, the Myb DNA binding domain. The Myb DNA binding domain was situated 24 AA from the N terminus of *A. thaliana* LHY and CCA1 and *O. sativa* LHY but 35 AA and 44 AA from the N terminus in *O. tauri* CCA1 and BbCCA1 respectively. The Myb DNA binding domain was 44 AA in length in all proteins compared (Figure 9 (a)).

Myb DNA binding domain sequences were highly conserved amongst BbCCA1, *A. thaliana* LHY and CCA1, *O. sativa* LHY and *O. tauri* CCA1 (Figure 9 (b)). The posterior probability score was maximum, *, for nearly all positions between short lower scoring regions at the beginning and end of the domain alignment where it was between 7 and 9. In two medial positions, in one of which there was a *B. braunii* substitution of an alanine for a serine residue, the posterior probability score was 8.

Following the Myb DNA binding domain in the sequences of CCA1 in *A. thaliana*, *O. sativa* and *O. tauri* there is a conserved signature; I/LPPPRPKRKPXXPYPQ/RK, which is also conserved in the sequence of BbCCA1.

(a)



(b)

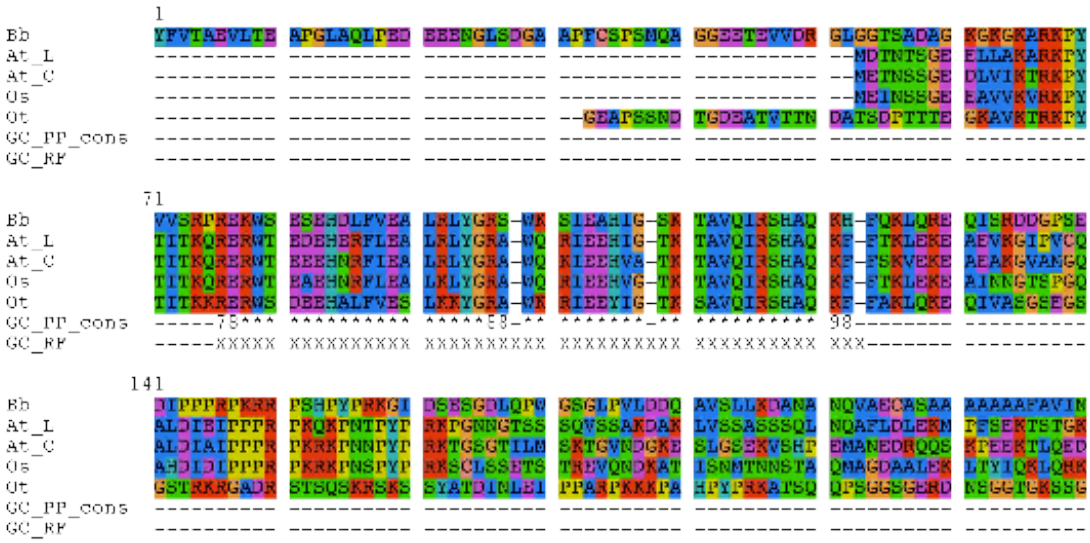


Figure 9 CCA1/LHY- like component domain architecture and alignments

(a) Schematic representation of BbCCA1 (Bb) domain structure compared with that of *A. thaliana* (At) LHY and CCA1, *O. sativa* (Os) LHY, and *O. tauri* (Ot) CCA1 and corresponding HMM alignments of the Myb DNA binding domain (b). On the 4th row is the posterior probability score of 0- 9*, where *= 0.95- 1.00. On the 5th row the HMM location is denoted by X.

B. braunii comp166095_c2_seq1 contained three RNA binding Kelch domains and a protein binding WW domain, as did CHLAMY1 subunit C1 (Figure 10). *B. braunii* comp166095_c2_seq1 was slightly longer (531 AA) than the *C. reinhardtii* counterpart (518 AA) but domain lengths were very similar; 59- 62 AA for the KH repeats and 28-29 AA for the WW domain in both organisms. Despite good sequence homology and domain architecture similarity between *B. braunii* comp166095_c2_seq1 and the *C. reinhardtii* clock element, CHLAMY1 (Figure 10), it was to the *O. tauri* clock and higher plant clock proteins that *B. braunii* transcripts were compared in subsequent analyses, as these models were better characterised for more informative comparison, they also represented both more basal and more complex evolutionary developments of plant circadian clocks respectively.

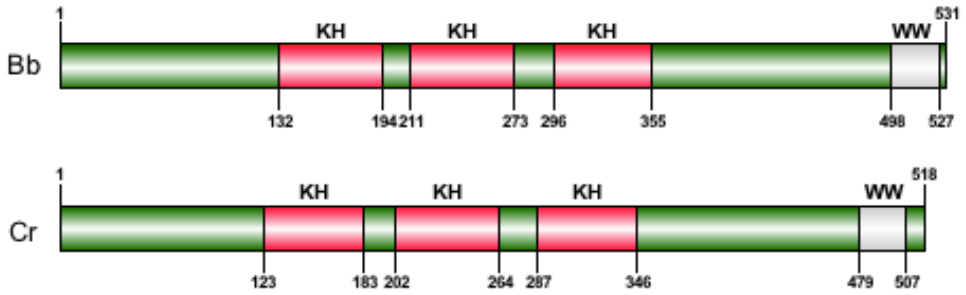


Figure 10 CHLAMY1 subunit C1 domain architecture and alignments
Schematic representation of *B. braunii* comp166095_c2_seq1 (Bb) domain structure compared with that of *C. reinhardtii* CHLAMY1 subunit C1. Kelch repeats are shown in red and the WW domain in grey.

4.3.2.3 Custom HMM generation and scan of Botryococcus braunii *ORFs*

A. thaliana TOC1, *A. thaliana* PRR3 and *O. tauri* TOC1 sequences were used to build the pseudo response regulator HMM. 222 *B. braunii* ORFs with domains matching the custom pseudo response regulator HMM were found by scanning the *B. braunii* HMM database, 134 of these had an E value ≤ $1e^{-05}$. The top 30 had E values between $2.80e^{-14}$ and $2.80e^{-58}$. There were three *B. braunii* ORFs that appeared as alternative sequences more than once in the top eight positions of the Top 30 table, these were comp56012_c0 seqs 1 and 2, comp169833_c4 seqs 9, 7 and 1 and lastly comp56565_c0 seqs 1 and 2.

B. braunii ORFs	E value
comp56012_c0_seq1_fr5	$2.80e^{-58}$
BbPRR_fr5	$4.50e^{-58}$
comp169833_c4_seq9_fr3	$2.60e^{-34}$
comp56565_c0_seq2_fr4	$1.30e^{-33}$
comp56565_c0_seq1_fr4	$1.30e^{-33}$
comp169833_c4_seq3_fr	$9.00e^{-33}$
comp169833_c4_seq7_fr3	$9.30e^{-33}$
comp168638_c2_seq1_fr3	$4.50e^{-31}$
comp156942_c1_seq1_fr5	$1.70e^{-29}$
comp111003_c0_seq1_fr1	$1.30e^{-26}$
comp47858_c0_seq1_fr2	$2.10e^{-23}$
comp160862_c0_seq2_fr3	$1.00e^{-19}$
comp144630_c0_seq1_fr3	$9.10e^{-19}$
comp169833_c4_seq2_fr3	$1.90e^{-18}$
comp169833_c4_seq4_fr3	$1.90e^{-18}$
comp153700_c0_seq1_fr1	$7.40e^{-18}$
comp140583_c0_seq1_fr6	$1.80e^{-17}$
comp168638_c3_seq1_fr2	$2.90e^{-17}$
comp160764_c0_seq1_fr2	$3.00e^{-17}$
comp168638_c3_seq2_fr2	$1.00e^{-16}$
comp108388_c0_seq1_fr3	$9.50e^{-16}$
comp297633_c0_seq1_fr1	$2.10e^{-15}$
comp1239458_c0_seq1_fr1 -	$3.70e^{-15}$
comp153591_c0_seq1_fr3	$5.20e^{-15}$
comp78703_c0_seq1_fr3	$5.50e^{-15}$
comp56199_c0_seq1_fr5	$1.70e^{-14}$
comp96719_c0_seq1_fr1	$2.80e^{-14}$

Table 4 Top 30 pseudo response regulator HMM ORFs

The left column lists the top 30 *B. braunii* ORFs matching a custom HMM generated from model clock pseudo response regulator protein sequences and their respective E values are shown in the right column.

109 *B. braunii* ORFs were identified by scanning the *B. braunii* HMM database using a custom HMM generated from *A. thaliana* LHY, *Oryza sativa* LHY and *O. tauri* CCA1. There were 30 ORFs with matching HMMs with E values ≤ $1e^{-05}$, the 12 smallest E values of these were alternative sequences of BbCCA1. The BbCCA1 ORFs had E values much smaller than the remaining 15 results; $2.30e^{-39}$ to $9.20E^{-40}$ compared to $1.40E^{-05}$ to $4.40E^{-13}$ respectively. The next smallest E value corresponded to comp149716_c0_seq1 ($1.50e^{-33}$).

B. braunii ORFs	E value
comp170985_c0_seq10_fr6	$9.20e^{-40}$
comp170985_c0_seq12_fr6	$9.20e^{-40}$
comp170985_c0_seq13_fr6	$9.20e^{-40}$
comp170985_c0_seq5_fr6	$9.20e^{-40}$
comp170985_c0_seq9_fr6	$1.40e^{-39}$
comp170985_c0_seq4_fr6	$1.40e^{-39}$
comp170985_c0_seq7_fr6	$1.70e^{-39}$
comp170985_c0_seq19_fr6	$1.70e^{-39}$
comp170985_c0_seq20_fr6	$1.90e^{-39}$
BbCCA1_fr6	$1.90e^{-39}$
comp170985_c0_seq16_fr6	$1.90e^{-39}$
comp170985_c0_seq17_fr6	$2.30e^{-39}$
comp149716_c0_seq1_fr1	$1.50e^{-33}$
comp58246_c0_seq1_fr5	$2.30e^{-29}$
comp56131_c0_seq1_fr5	$4.40e^{-13}$
comp173589_c8_seq1_fr5	$2.40e^{-12}$
comp792752_c0_seq1_fr5	$3.20e^{-12}$
comp162544_c0_seq1_fr4	$5.90e^{-11}$
comp160872_c2_seq1_fr1	$7.30e^{-11}$
comp145220_c0_seq1_fr1	$2.00e^{-09}$
comp157845_c0_seq1_fr3	$4.10e^{-09}$
comp163242_c0_seq4_fr1	$2.20e^{-07}$
comp149649_c2_seq2_fr4	$5.40e^{-07}$
comp160564_c0_seq1_fr1	$3.10e^{-06}$
comp162544_c0_seq4_fr4	$8.40e^{-06}$
comp162544_c0_seq5_fr4	$8.40e^{-06}$
comp171017_c18_seq1_fr5	$1.40e^{-05}$

Table 5 Top 30 myb DNA binding HMM ORFs

The left column lists *B. braunii* ORFs matching a custom HMM generated from Myb DNA binding type model clock proteins and on the right are shown their corresponding E values.

4.3.2.4 Keyword domain search of Botryococcus braunii *ORFs*

After the RRR and CCT domains from the pseudo response regulators of model clocks were identified (4.3.2.1 Domain and motif identification) the domain names were used as keywords to search the *B. braunii* transcriptome HMM database. There were 18 transcripts that had detectable CCT domains and 273 transcripts that had detectable RRR domains. Only two of the 18 transcripts that contained CCT domains also contained an RRR domain, these were alternative sequences of the same predicted ORF; comp56012_c0_seq1 and BbPRR. The BLAST top hit of comp56012_c0_seq1 and seq2 (BbPRR) was the same; both with *Coccomyxa subellipsoidea* C-169 hypothetical protein, E value = $9e^{-32}$.

The keyword search for 'Myb' domains in the *B. braunii* transcriptome HMM database yielded 272 ORFs containing the Myb DNA binding domain, 12 alternative sequences for comp170985 (BbCCA1) were in this list, with a small range in E values ($6.9e^{-13}$ to $2.8e^{-13}$). The ORFs with E values ≤ $1e^{-10}$ were searched for in the BLAST top hits output of the *B. braunii* transcriptome, none were found. However, a search of 'Myb' in the BLAST output yielded 16 results although none of these were from within the *Viridiplantae*. A search for comp170985 sequences in the BLAST annotation of the *B. braunii* transcriptome showed that alternative sequences 4, 7, 9, 11, 12, 13, 16, 17, 19 and 20 of comp170985 all resulted in a BLAST hit with a predicted protein of *Ostreococcus lucimarinus* CCE9901 (E value $1e^{-18}$).

4.3.2.5 The Botryococcus braunii *clock model*

Using cross- referenced data from studies of the above organisms (Murakami *et al*., 2006; Pokhilko *et al*., 2012; Troein *et al*., 2011), two hypothesised clock models were generated, one to represent that of higher plants (Figure 2) and one for the simpler mechanism of green algae (Figure 3). Based upon the results obtained within this chapter, a model of the *B. braunii* clock was generated (Figure 11).

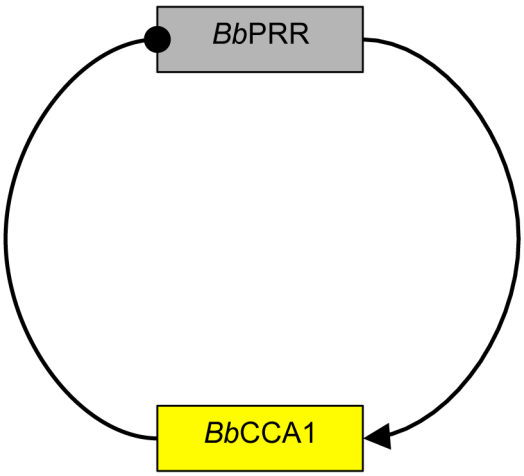


Figure 11 *B. braunii* clock model
The proteins of the proposed simple clock mechanism of *B. braunii* are shown with the morning component, an CCA1- like protein in yellow and the evening component, a pseudo response regulator- like protein in grey. Transcriptional regulation is shown by solid black lines terminating with a solid circle for down- regulation and a solid arrowhead for up- regulation.

4.4 Discussion

4.4.1 Identification of clock protein homologs by sequence

In this chapter a wide net was initially cast to identify circadian clock genes in the *B. braunii* transcriptome, using sequence homology searches for circadian clock components. BLAST search results revealed five *B. braunii* sequence homologs to clock components from higher plants and green algae.

Protein sequences were used for the BLAST search because this method circumvented the issue of frame shift, codon usage disparities and other ambiguities that may cause failure to detect certain open reading frames.

Five PRR genes previously characterised in the short- day plant *Oryza sativa,* are similar to corresponding *A. thaliana* PRRs (Murakami *et al*., 2006). However, in both *A. thaliana* and *O. sativa* highly divergent amino acid sequences in the regions between the CCT and RRR domains of pseudo response regulators render each component; 1 (TOC1), 3, 5, 7 and 9 unique (Matsushika *et al*., 2000; Mizuno & Nakamichi, 2005). BLAST searches of *A. thaliana* TOC1 and PRRs 3, 5, 7 and 9 against the *B. braunii* ORF database resulted in the recurrence of one *B. braunii* protein, comp56012c0seq2 (Table 2), indicating that instead of a set of PRRs, as in the higher plants, only one may be present in the alga, as is the case in *O. tauri* (Corellou *et al*., 2009). Based on the protein sequence data alone it is difficult to infer which specific model clock PRR *B. braunii* BbPRR corresponds to. Whilst *O. sativa* PRR1 was identifiable as a homologue to TOC1 in *A. thaliana*, *O. sativa* PRRs 5 and 9 were hard to distinguish between, and the same with PRR7 and 3 (Murakami *et al*., 2006). In *O. tauri* putative homologs to core clock genes in higher plants were unidentifiable based on sequence homology but instead CCA1/LHY and PRR1 counterparts to higher plant clock components were discovered based on conserved domains (Corellou *et al*., 2009). Therefore a similar strategy was employed to mine the *B. braunii* transcriptome for circadian clock components.

Positive BLAST results with the alternative and conserved plant core clock components; TOC1, the PRRs and CCA1/LHY supported by their associated peripheral components ELF3, ELF4 and LUX directed analyses away from the *C. reinhardtii* clock mechanism, involving CHLAMY1. ELF3 and 4 are found exclusively in land plants thus far and so a point of interest for future work may be to further investigate the function of *B. braunii* sequence homologs found for these proteins, as this is a novel discovery (Locke *et al*., 2006).

Ultimately the sequence homology results and initial domain comparisons were sufficient to direct the focus of subsequent analysis on to the pseudo response regulators and CCA1/LHY- type proteins.

4.4.2 Identification of clock protein homologs by functional domain

4.4.2.1 Domain and motif identification

The domain architecture in the *B. braunii* homologues to model clock components was used to narrow the sequence homology results down, resulting in two predicted clock components; one counterpart to the PRR family, BbPRR and one representative of the LHY/ CCA1 type components of *Viridiplantae*, BbCCA1.

Interestingly comp112899_c0_seq2, the *B. braunii* top hit against *O. tauri* PRR1 was different to that arising from *A. thaliana* (BbPRR, E value of $1e^{-24}$) and had a lower E value of $5.00e^{-68}$. However the RRR and CCT domains were not conserved, instead comp112899_c0_seq2 contained an acyl-coA domain and Kelch repeats. Kelch repeats are conserved motifs throughout the Eukaryotes with various functions but their primary purpose in plants seems to be substrate recruitment in F- box proteins. F-box proteins recruit proteins for ubiquitination and are thus involved in signal transduction (Andrade *et al*., 2001; Kepinski & Leyser, 2005). However in place of a detectable N-terminal F- box domain, comp112899_c0_seq2 has an acyl- CoA binding domain. *A. thaliana* proteins involved in lipid metabolism via the transport of oleoyl- A to the ER have been identified containing an N- terminal acyl-CoA domain and C-terminal Kelch repeats as observed in comp112899 indicating a similar role for this protein in *B. braunii* (Leung *et al*., 2004).

4.4.2.2 Domain architecture comparison with model clock components

The inter-domain regions of the model clock proteins from *A. thaliana*, *O. sativa* and *O. tauri* showed low sequence conservation with *B. braunii* and with each other. HMM alignment was used to align the conserved domains to downloaded HMMs of model clock components, so that sequence similarity in conserved domains could be assessed.

Profile HMMs were used because of their efficacy in identification of analogous proteins with divergent sequences. Probabilistic characterisation of local areas of similarity, regardless of interrupting areas of dissimilarity that were likely irrelevant to function by incorporating weighted schemes and priors allowed the development of response regulator and CCA1/LHY protein domain models.

Domain visualisation of the pseudo- response regulators and CCA1/LHY proteins of *A. thaliana*, *O. sativa* and *O. tauri* revealed that the domains of core plant clock components were conserved and positioned in the same way within *B. braunii* BbPRR and BbCCA1 (Figure 4 (a), Figure 5 (a), Figure 6 (a), Figure 7 (a), Figure 8 (a), Figure 9 (a)). Furthermore, HMM alignments confirmed that sequence was highly conserved within the domains of these proteins from *B. braunii* and the model organisms compared (Figure 4 (b), Figure 5 (b), Figure 6 (b), Figure 7 (b), Figure 8 (b), Figure 9 (b)).

The PRRs are structurally similar to authentic response regulators (ARRs), which are conserved throughout prokaryotes and eukaryotes (except vertebrates). Response regulators (RRs) are part of a two- component signal transduction pathway, in which RRs act as phosphate acceptors from a histidine kinase (HK) donor upon a specific stimulus, imparting control (usually as a transcription factor) over a particular cellular event (D'Agostino & Kieber, 1999; Hwang, 2002; Mizuno, 1997; Mizuno & Mizushima, 1990). It is proposed that the *Viridiplantae* evolved their own specialist signal transduction components (the PRRs) for incorporation into light signaling pathways and the circadian clock (Mizuno & Nakamichi, 2005).

O. sativa and *A. thaliana* PRRs share a unique glutamate (E) residue in the middle of the RRR domain, a trait suggested to distinguish pseudo response regulators from authentic response regulators, which have a phosphate- accepting aspartate (D) residue at this position (Murakami *et al*., 2003). Interestingly, *O. tauri* PRR1 and BbPRR both have an aspartate in the position of the higher plant substitution of aspartate for glutamate, likening them to ARRs (Figure 4). However, known plant ARRs contain a C-terminal GARP DNA binding domain, as opposed to the CCT domains present in *O. tauri* PRR1 and BbPRR (Hosoda, 2002; Putterill *et al*., 1995). The CCT domain is named after the proteins within which it was first identified; CONSTANS, the CONSTANS-like family of transcription factors and TOC1, which are all involved in the higher plant clock (Strayer *et al*., 2000).

The conserved I/LPPPRPKRKPXXPYPQ/RK signature of CCA1 and LHY homologues in *A. thaliana*, *O. sativa* and *O. tauri* (Corellou *et al*., 2009) was observed in BbCCA1, similarly positioned a short distance towards the C- terminus after the Myb DNA binding domain (Figure 9). The discovery of this signature was fortuitous as the only conserved domain for CCA1/LHY type clock protein identification is the Myb DNA binding domain, common to many plant transcription factors (Romero *et al*., 1998). It was noted that the accession for *O. tauri* CCA1 protein sequence given by Corellou *et al* is incorrect; the correct accession is given in Table 1.

K Homology (KH)- repeat containing proteins such as CHLAMY1 and *B. braunii* comp166095_c2_seq1 are nucleic acid- binding and are conserved throughout the Eukaryotes. The *A. thaliana* genome encodes 26 different KH domain containing proteins, which are RNA- binding and one of the most prevalent types of transcriptional regulatory protein (Lordovic & Barta, 2002). *B. braunii* comp166095 also shared a C-terminal WW domain in a similar relative position as CHLAMY1 (Figure 10). WW domains are highly conserved in the Eukaryotes and bind to particular proline- rich motifs in target proteins. The moniker WW refers to the two conserved tryptophan residues spaced 20- 23 positions apart within the domain (Hesselberth *et al*., 2006; Macias *et al*., 2000).

Algal clock components were first identified in *C. reinhardtii* (Iliev *et al*., 2006; Matsuo *et al*., 2008; Schmidt, 2006). However, until recently the *C. reinhardtii* core oscillator was unresolved although an RNA-binding protein complex, CHLAMY1, and a serine/ threonine casein kinase (CK1) were recognised to be involved. A definite and precise mode of action for CHLAMY1 within the core oscillator was not decided upon. Adding to the complexity was the action of CHLAMY1 as a translational repressor of expression, binding to the 3' un- translated regions of mRNAs, particularly those involved in nitrogen and carbon metabolism (Matsuo & Ishiura, 2011; Mittag, 1996). Seemingly CHLAMY1 not only has a role in the central clock of *C. reinhardtii* but also functions as an output of the clock (Iliev *et al*., 2006).

CK1 activity has multivariate implications in *C. reinhardtii*, including affects on circadian rhythmicity (Schmidt 2006). Furthermore casein kinase proteins are conserved throughout the Eukaryotes in diverse and numerous pathways, including input and regulatory pathways of the circadian clock in higher plants (Gross & Anderson, 1998; Hori *et al*., 2000; Park, 2012). Identification of a CK1 homolog in *B. braunii* would therefore have little impact in terms of elucidating components of a circadian molecular oscillator, particularly whilst an optimised genetic transformation toolkit is lacking, which would allow demonstration of circadian function, e.g. by way of a CK1 knockout experiment.

CHLAMY1 *c1* mRNA expression is constant throughout the diurnal cycle and therefore comparison of expression profiles, as performed with other cycling homologs in the forthcoming chapter, would not yield any further information. However, CHLAMY1 is a heteromeric protein, consisting of subunits containing three RNA recognition motifs (C3) and another containing three KH motifs (C1) (Zhao *et al*., 2004). Whilst the *c1* mRNA remains constant in expression over the diurnal period, *c3* expression oscillates (Kucho *et al*., 2005). Therefore comparative mRNA expression

profiling with a CHLAMY1 C3 subunit homolog in *B. braunii* has potential to elucidate core circadian components where comparison with *c1* has not.

Whilst the sequence homologues of *A. thaliana* LUX and *C. reinhardtii* CHLAMY1, were ultimately discounted from the work described in this chapter they may warrant further investigation although this was either beyond the scope or time constraints of this project.

4.4.2.3 Custom HMM generation and scan of Botryococcus braunii *ORFs*

The results of a scan of the *B. braunii* ORF database with an HMM generated from model PRR proteins supported the findings of the conserved domain comparisons and sequence homology with proteins from other plants. Profiles of alternative sequences of BbPRR matched the PRR HMM generated and were the results with the lowest E values (Table 4). Interestingly the sequence homolog of *A. thaliana* LUX also featured in the top twenty results of the PRR HMM scan although the presence of the RRR domain (absent in *A. thaliana* LUX) in *B. braunii* comp56565_c0_seq2 could account for this. BbCCA1 also had a profile matching that of the CCA1/LHY HMM, with the alternative sequences of this protein comprising the results with the lowest E values of annotation.

4.4.2.4 Keyword domain search of Botryococcus braunii *ORFs*

Two proteins, both alternative sequences of BbPRR, were annotated with a detectable RRR and a CCT domain by the Pfam scan. Due to the single conserved Myb DNA binding domain of the CCA1/LHY type components, searching for proteins annotated by Pfam to have an Myb DNA binding domain yielded a large number of results. These results are unsurprising as Myb proteins are a large family of transcription factors with involvement in diverse processes from regulation of secondary metabolism to cellular morphogenesis (Jin & Martin, 1999). However, twelve alternative sequences of BbCCA1 were the results with the lowest E values of proteins annotated with an Myb DNA binding domain.

The assignment of Pfam domains and subsequent searches using HMMs of model clock proteins and keywords based on conserved domain names both yield results lending support to the proposal of BbCCA1 as a CCA1/LHY homolog and BbPRR as a PRR family protein. BLAST results for these proteins provided reassurance that they were indeed of algal origin, as opposed to contaminating sequences or sequencing errors. Many other proteins in the *B. braunii* database were also annotated with similar domain architecture to model clock components and may be worth investigating in future searches for circadian components of *B. braunii*.

4.4.3 The Botryococcus braunii clock model

Comparative studies based on the complete *O. sativa* genome revealed a likely ortholog of each *A. thaliana* circadian component, indicating evolutionary conservation of a molecular basis for circadian rhythm between dicotyledonous (e.g. *A. thaliana*) and monocotyledonous (e.g. *O. sativa*) plants (Murakami *et al*., 2003). Conserved domains and sequences demonstrated by sequence alignments in this chapter confirm a high degree of similarity between *A. thaliana* and *O. sativa* clock architecture, whilst *B. braunii* and *O. tauri* similarity only extends as far as a single PRR and a single CCA1/LHY type component. However it is proposed that gene expansion resulting in the collection of PRR components in the higher plant circadian clock occurred via polyploidy events in the divergence of the monocots and eudicots, therefore it is unexpected that multiple PRR components will be present in *B. braunii*, which represents a group of organisms that were ancestral to the higher plants (Takata *et al*., 2010).

A robust and flexible two- component oscillator comprising a PRR family protein and a CCA1/LHY type protein is maintained in *O. tauri* (Thommen *et al*., 2012). Expression profiles of *O. tauri* TOC1 and CCA1 were accurately recreated using a mathematical model of a two- component transcriptional/ translational feedback loop (Morant *et al*., 2010; Thommen *et al*., 2010). A simpler, single-loop version of that of higher land plants is also suggested in the moss *Physcomitrella patens*. *P. patens,* is a basal land plant, phylogenetically closer to a shared ancestor with *B. braunii* than higher land plants such as *A. thaliana*. In this chapter one homolog to the CCA1/LHY type clock components and a PRR protein are identified in *B. braunii*; comp170985 (BbPRR) and comp56012 (BBCCA1) respectively. Contrastingly, *P. patens* has two homologs of the CCA1/LHY- type clock components; *Pp*CCA1a and *Pp*CCA1b and a set of four PRRs- *Pp*PRR1-4. However *P. patens* homologs to TOC1 and other higher plant clock components, GI and ZTL have not been established (Holm *et al*., 2010).

In a phylogenetic analysis of the RRR domains of PRRs in a previous study, three distinct clusters were formed; one containing the higher plant TOC1s, another with the remaining PRR family proteins and the third comprised of proposed algal TOC1s. None of the PRRs of *P. patens* or the club moss, *Selaginella moellendorffi* clustered near either algal TOC1s or those of the higher plants, suggesting no ortholog of TOC1 in these basal land plants. The algal TOC1s seemed to be divergent from those of higher plants and the rest of the PRRs, indicating they are in fact a sister lineage as opposed to true TOC1 orthologs. Additionally, the conserved aspartate (instead of glutamate in higher plants) in the RRR domains of BbPRR and *O. tauri* TOC1 implies recent divergence from ARRs. Furthermore, the identification of only one

candidate PRR in both *B. braunii* and *O. tauri* suggests that selection pressures on vascular plants led to duplication of the PRRs (Holm *et al*., 2010). Data from previous studies and that of this chapter imply a common ancestor of PRR genes was present in the Chlorophyta prior to the emergence of higher plants.

Other studies of circadian topology in green algae concurr with that of this investigation, with discovery of homologs to the higher plant clock limited to a single representative of the PRRs and CCA1/LHY type components. The output of this chapter is a proposed *B. braunii* clock model akin to that of *O. tauri*, comprising just two proteins; BbPRR- a PRR homolog and BbCCA1- a CCA1/LHY homolog (Figure 11). However for further confirmation of involvement in the circadian clock for *B. braunii* predicted PRR and CCA1 proteins, expression pattern analysis was necessary.

The investigation and successful identification of clock components imparts confidence in the suitability of the *B. braunii* transcriptome generated in this study to be used as the basis for a deeper exploration and therefore, understanding, of the molecular physiology of *B. braunii*.

Chapter 4 Summary

In this chapter, the annotated *B. braunii* transcriptome was mined for circadian clock components by performing a TBLASTX search using existing and model clock proteins as queries. This analysis was performed to assess whether the transcriptome data, acquired in both diurnal and circadian regimes, could be used to characterise a molecular circuit that was previously unknown, but in all probability present in the algae. Positive BLAST hits were further characterised by identification of conserved domains and comparison with those of model clock components. In order to investigate sequence conservation within functional domains, Hidden Markov Models were generated and used for sequence alignment. A model of the *B. braunii* circadian clock was constructed and resembles more closely that of the green alga *Ostreococcus taurii* than that of *Arabidopsis thaliana*, indicating the high level of clock control systems that evolved in angiosperms, compared to the chlorophyta.

CHAPTER 5

ANALYSIS OF TIME- DEPENDENT DIFFERENTIAL EXPRESSION IN THE *BOTRYOCOCCUS BRAUNII* TRANSCRIPTOME

5.1 Introduction

Core components of circadian clocks in model organisms have clear oscillatory patterns in their transcript expression. For example, in *A. thaliana* TOC1 expression builds throughout the afternoon and peaks shortly before dusk, falling steeply thereafter and then accumulating again a few hours after dawn, creating daily waves of expression (Hsu *et al*., 2013; Matsushika *et al*., 2000). By definition, the expression of circadian clock components maintains an oscillatory pattern even when placed under constant conditions, although the amplitude of the wave may be damped.

Aim 1 of this chapter is to confirm the expression patterns of predicted *B. braunii* clock components generated from mRNA transcripts in Chapter 3 with model clock component expression under two different diel regimes; a 12L/ 12D diurnal photoperiod and constant light.

Aim 2 of this chapter is to gain a holistic view of pathways and processes that are under circadian or diel control in *B. braunii* by comparing the expression of transcripts under the two photoperiodic conditions throughout a time series of 28 hours. Statistical analysis of differential expression (DE) between conditions and timepoints will be used to address whether transcript concentration differs between samples and if these differences can be attributed to the condition or timepoint. Lastly this chapter aims to elucidate networks of co-expression of pathways and genes in *B. braunii* by cluster analysis.

In contrast to the dataset in previous chapters, in this chapter sequence data is retained within individual timepoints and is the focus of analyses in this chapter, although the *B. braunii* reference transcriptome generated in Chapter 3 from the collective sequence data of all samples is used for alignment (Figure 1). Sequence reads from within each timepoint are aligned to the *B. braunii* transcriptome and the abundance of each transcript estimated from read counts. Using this data, expression of predicted transcripts is compared throughout the time series, statistical significance

of DE is inferred and for target genes of interest, such as predicted clock components and those involved in terpenoid synthesis, patterns of expression are plotted and comparisons are drawn with expression patterns of characterised model genes.

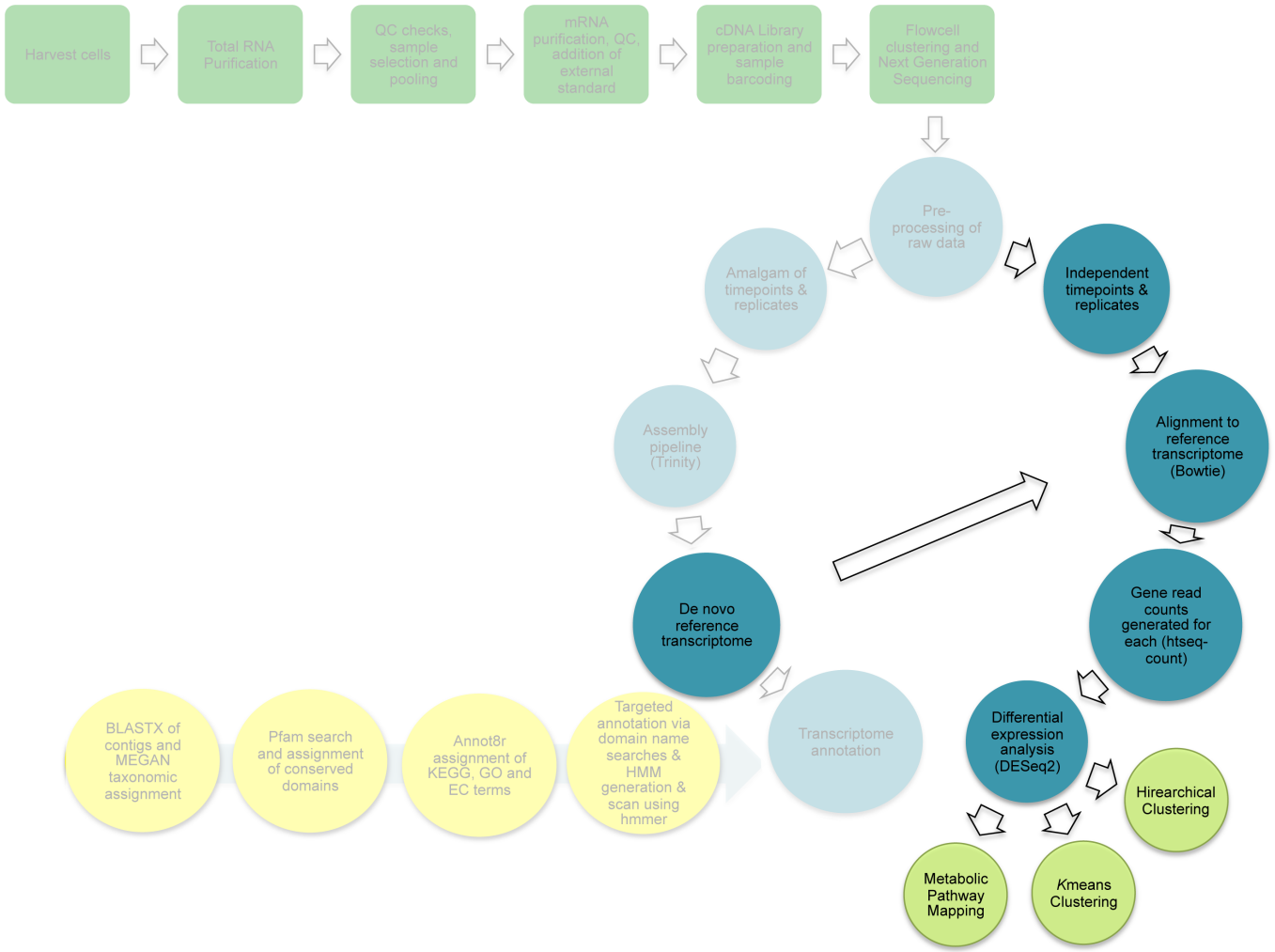


Figure 1 Experimental and bioinformatic methods

Practical laboratory methods (green) through to bioinformatic analysis (blue) used to investigate the transcriptomics of *B. braunii* over a 28-hour time series under two different photo-regimes. This chapter describes the differential expression and co-expression analysis of the *Botryococcus braunii* transcriptome (teal blue and lime green) sequence. Faded out components are addressed in Chapters 3 and 4.

5.2 Materials and Methods

5.2.1 Alignment of timepoint reads to reference transcriptome assembly

The bowtie-build tool from the Bowtie aligner toolkit version 4.8.0 20130211 (Langmead *et al*., 2009), was used to generate a compressed sequence index of the *B. braunii* transcriptome (generated in Chapter 3) using the Burrows Wheeler data compression transform (Burrows & Wheeler, 1994). The Bowtie alignment tool was then used to align reads from independent time- points and replicates to the indexed reference transcriptome. 16 parallel search threads were used per alignment, generating a Sequence Alignment/ Map file (SAM) for each sample. SAM files are a generic alignment format file in which read alignments against reference sequences are stored; multiple downstream alignment processing pipelines are supported by this format (Li *et al*., 2009).

5.2.2 Generation of count data using HTSeq

Using a Bioperl (Stajich, 2002) script available from Sourceforge (http://sourceforge.net/p/gmod/mailman/message/31345898/) a general feature file was created from the *B. braunii* transcriptome, giving coordinates for features within, in this case, annotated transcripts of the *B. braunii* transcriptome. The number of reads overlapping with transcripts in the *B. braunii* transcriptome were then counted using the Python application, HTSeq-count, available as part of the HTSeq software, (version 0.5.4p5) (Anders *et al*., 2014). The sequence alignment files generated in section 5.2.1 were provided to HTSeq-count and reads mapping within the coordinates specified for transcripts by the general feature format file of the transcriptome were counted for each sample. All other parameters were ran as ascribed by the default for HTSeq-count.

A custom BioPerl script (htseqcount2deseq.pl, Appendix disc) was used to merge the HTSeq-count output files into one count matrix- a single tab delimited file to be used as input for DE analysis by DESeq2. The count matrix, *K*, of data created contained one row for each transcript, *i*, and one column for each sample, *j*. Separate counts files for LD and LL samples were also generated manually by opening the total counts file as a spreadsheet and selecting columns to copy into the separate counts files for the different conditions.

5.2.3 Differential expression analysis by DESeq2

The count matrix was imported into DESeq2, version 1.4.5 (Love *et al*., 2014), an R package (Team) available in the Bioconductor software suite (http://www.bioconductor.org/install/). DESeq2 uses a negative binomial distribution to model counts and infer statistical significance to DE of RNA-Seq data. The described method was executed using a script modified from the vignette available with the DESeq2 software (Appendix disc).

In order to check for batch effects in the dataset, count data was ordered by not only timepoint and condition (i.e. LD or LL) but also sequencing lane. Basemeans, log fold change and adjusted p values were calculated for each pair- wise comparison between timepoints in LD and LL, for example 0LDP1 was compared to 0LLP1 and so on. Log2 fold change for each transcript was given by the log2 of counts in one sample divided by normalised counts in the compared sample. An MA plot of log fold changes of transcripts in LD and LL against the means of normalised counts was drawn.

Normalisation of counts was performed by the division of transcript expression values by a Size Factor calculated for each sample. Size Factors were given by division of every transcript expression value within a sample by the geometric mean expression of that transcript across all samples resulting in a reference expression value for each transcript within a sample. The median reference expression value equated to the Size Factor for that transcript.

Transcripts that had expression that was extreme, in one direction or another in only one sample were removed from the dataset by estimation of the contribution of each sample to the log fold change of a gene, calculated by Cooks Distance followed by exclusion of the gene from DE analysis if they had an extreme affect. The Cooks Distance cutoff used for removal of outlying transcripts was the default of 0.99 quantile of the F (p*, m - p*) distribution, where *p* is the number of model parameters (the negative binomial model) and *m* is the number of samples. Maximum likelihood estimate was used to calculate Cooks Distance (Love *et al*., 2014).

The dataset was subsequently separated into the two conditions; LD and LL. The likelihood ratio test was used to test each sample for DE using the 0LD and 0LL timepoints as a reference for each condition respectively.Adjusted p value was used as a confidence threshold for temporal DE detection. Transcripts with temporal DE (padj≤0.01) from the LD and LL datasets were binned into separate lists in text format files, the same was performed for transcripts with no temporal DE (padj>=0.01) using R data sorting and extraction commands (scripts LL_LRD.R and LD_LRD.R, Appendix disc). Lists were cross- referenced using Venn diagrams generated by the Venny online interactive tool () to create subsets of transcripts (Table 1). The normalised

count data for temporally differentially expressed transcripts was exported into MATLAB for subsequent analysis and visualization (Mathworks).

5.2.4 Removal of outliers

5.2.4.1 Candida glabrata contamination

The initial MA plot revealed unusual and unexplained patterns in expression of a cluster of transcripts with log fold change of -4 and less (Figure 2). The outliers were investigated by extracting all transcripts with a log2 fold change of -4 or less from the DESeq2 data set and cross referencing them against the BLAST annotation data generated for the entire *B. braunii* transcriptome (see Chapter 3). The BLAST search process was still in process but of the 9,015 transcripts with log2 fold change -4 or less, 2,776 had already been queried by BLAST, 80% of which had a significant (E value $<1e^{-05}$) hit against *Candida glabrata* strain cbs138.

B. braunii transcriptome transcripts were searched against an index of the *C. glabrata* genome (downloaded from NCBI, project code PRJNA12376) using gmap (see Chapter 3). The protein sequences of the transcripts successfully aligned by gmap were imported into an Excel spreadsheet and an alignment quality value calculated by division of the number of mismatches by length. Therefore small alignment quality value indicated a low number of mismatches to alignment length and hence a good alignment. The results were sorted in order of ascending alignment quality and sequentially searched against the NCBI nr database. Based on these results an alignment quality cutoff value of 0.016 (the cutoff score of the lines of data in which BLAST results were starting to include *Viridiplantae*) was decided upon. Names of transcripts with alignment quality values above this were used to create a text file list. Using the list as a search term, the command line tool grep was used to remove from the count matrix the *C. glabrata* transcripts. DE analysis was recommenced from import of count matrix onwards.

5.2.4.2 Screening for anomalous patterns

Hierarchical cluster analysis and generation of a cluster tree of samples was performed in R, by calculation of a Spearman's rank test to measure similarity between the expression profiles of all samples (see Appendix disc- Charlotte_correlation.R). R commands were also used to draw a boxplot of reads from each sample, that were counted overlapping with transcripts of the *B. braunii* transcriptome by HTSeq-count (Appendix disc).

5.2.4.3 Removal of samples with low read alignment number

Samples with a low alignment number that were highlighted by hierarchical clustering and plotting of reads overlapping transcripts per sample (section 5.2.4.2) were removed from the counts files generated by HTSeq-count by deleting the sample columns from the tabulated text file prior to re-commencing DE analysis.

5.2.5 Targeted expression analysis of genes of interest

5.2.5.1 DE of predicted clock components and terpenoid pathway genes

Chapter 4 provided a short-list of *B. braunii* predicted core clock component transcript numbers. To determine whether the predicted clock components were differentially expressed in a temporal manner and under which photo-regime this occurred, their transcript numbers were searched for within the subsets of transcripts (Table) using the Linux command line tool grep.

The assignment of Enzyme Commission numbers (EC), Kyoto Encyclopedia of Genes and Genomes (KEGG) identifiers and Gene Ontology (GO) terms to transcripts in Chapter 4 allowed the mapping of *B.braunii* transcripts to the KEGG terpenoid and sesquiterpenoid reference pathways. Initially, the phylogenies of transcripts that were mapped to the pathways of interest were confirmed to be *Viridiplantae* by searching against the BLAST annotation of the *B. braunii* transcriptome (section 3.2.4.1 BLAST sequence homology). However, the large presence of non- *Viridiplantae* BLAST annotations in the dataset made selecting for the *B. braunii* laborious. Subsequently, the EC and KEGG annotations of the *B. braunii* transcriptome were filtered using MEGAN (Section 3.3.3.3 KEGG pathway analysis), to include only transcripts annotated by BLAST to be within the *Viridiplantae*.

Datasets compared	Transcript list attributes
DE in LD vs DE in LL	Temporal DE under both photo-regimes
No DE in LD vs No DE in LL	No temporal DE in either photo-regime
DE in LD vs No DE in LL	Temporal DE in LD but not LL
No DE in LD vs DE in LL	Temporal DE in LL but not LD

Table 1 Subsets of transcripts
Lists of transcripts with temporal DE (padj≤ 0.01) (DE) or no temporal DE (padj≥ 0.01) (noDE) under either LD or LL conditions were compared in different pairwise combinations to generate Venn diagrams, outputting lists of transcripts common to both compared lists.

5.2.5.2 Expression patterns of predicted clock genes

The counts for replicates from each timepoint for putative *B. braunii* clock transcripts were extracted from the HTSeq-count output, mean values and standard deviation were calculated and mean counts were plotted against time.

5.2.6 Cluster analysis of expression

5.2.6.1 Scree plots and K-means

Using a MATLAB script (scree_plot.R, Appendix disc) the within groups sum of squared variance of transcript expression for cluster solutions of 1 to 15 was plotted comparing samples within their respective conditions, LD and LL. The *K*-means cluster solution was determined by the lowest cluster solution representative of the majority of variance in expression profile. K-Means clustering was performed using only transcripts with padj≤ 0.001 as the computation and visualisation of all possible pairwise comparisons between expression profiles of transcripts using a cutoff of padj≤ 0.01 was unfeasible.

5.2.6.2 Pearson correlation

R commands were used to calculate the Pearson correlation coefficient to measure similarity between gene profiles in a pairwise fashion, assigning values between -1 (perfect anti-correlation) and +1 (perfect correlation). To perform hierarchical clustering the Pearson correlation coefficients were subtracted from 1 to give a distance metric:

$$Dist_{PEARSON} = 1 - cor_{PEARSON}$$

Pearson distance metrics varying from 0 (perfect correlation) to 2 (perfect anti-correlation) were then used to iteratively group transcripts with the smallest distance metrics together forming clusters. After the initial clusters had been established, distance between clusters was calculated by the average linking method, which calculates the mean of all the distances between the transcripts in the first cluster and the transcripts in the second cluster, rearranging until for each transcript the closest mean solution is found (Quackenbush, 2001).

5.3 Results

5.3.1 Alignment of timepoint reads to reference transcriptome assembly

HTSeq-count counted 153,959,202 reads mapped transcripts in the *B. braunii* transcriptome from the LD dataset and 183,882,618 from the LL dataset. For the frequency of mapped reads from each timepoint individually see Table 2 in Appendix.

5.3.2 Outliers

Preliminary data plots revealed Samples 0LDP2, 20LDP2, 24LDP2, 16LDP3 and 12LLP2 as outliers, which were subsequently removed from the counts files for DE analysis (5.2.4.3 Removal of samples with low read alignment number). Samples 0LDP2, 20LDP2 and 24LDP2 had very low reads aligned to transcripts (Figure 19 in Appendix). Hierarchical clustering of expression data showed that samples 12LLP2 and 16LDP3 were atypical to the rest of the data and in further investigation both showed bias in a plastid gene and had very low expression in genes that were highly expressed in all other samples (Figure 21 in Appendix). Furthermore the number aligned reads aligned to *B. braunii* transcripts was very low for samples 12LLP2 and 16LDP3 (315, 982 and 244, 388 respectively) compared to the rest of the samples (Table 2 in Appendix).

5.3.3 DE between conditions

In a plot of log2 fold changes between timepoints from LL and LD conditions transcripts with an adjusted p value ≤ 0.1 (default parameter) were visualised in red (Figure 2). Significant DE was detectable in transcripts with mean normalised counts of as low as ~0.07. However, log2 fold change for transcripts with low normalised counts was required to be large, ranging from down- regulation of 4 to up- regulation of 2. There was a large and dense cluster of transcripts with average counts of between ~1 and one hundred and down- regulated log2 fold change of -4 and less. 16,835 transcripts originating from *C. glabrata* were removed from analysis after investigation of this cluster (section 5.2.4.1 Candida glabrata contamination). Subsequent to the removal of *C. glabrata* contamination, 287,421 transcripts remained within the dataset for DE analysis.

Only fold changes of 2 or more were deemed significant for transcripts with average counts above one hundred, however there are few in this category. Shrinkage of dispersion can be seen on the left side of the graph where transcripts with low counts but high dispersion are moderated to a greater extent than others, by a default

DESeq2 zero- centred normal prior. There were few differentially expressed transcripts with average counts below 1.

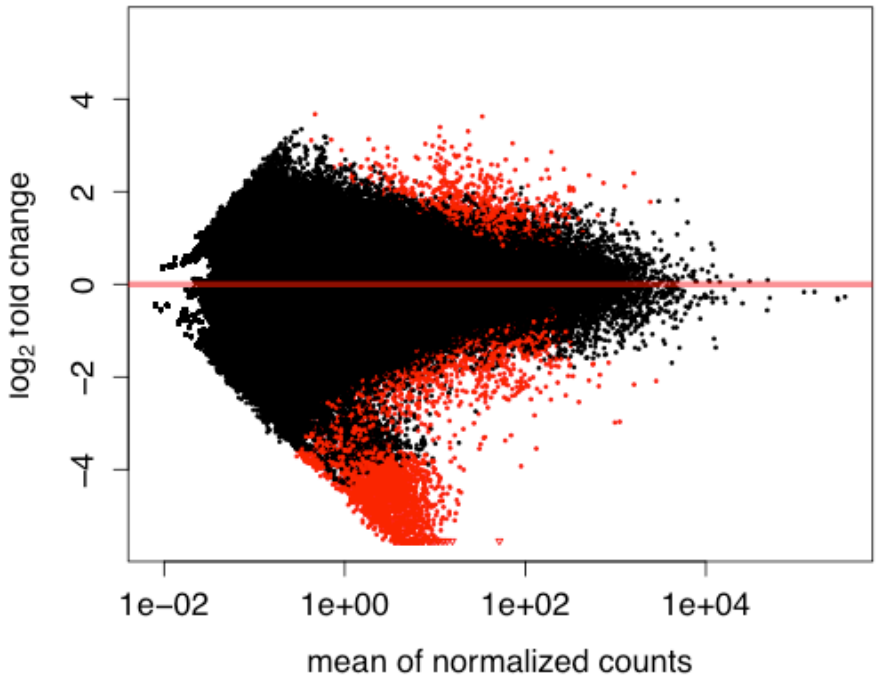


Figure 2 Expression dispersion of transcripts in LD and LL conditions
Average expression over all samples (χ axis) for all transcripts (dots) is plotted against $\log_2$ fold change (γ axis) between LL and LL conditions. Significantly differentially expressed transcripts (padj≤ 0.01) are shown in red. Tapering of points towards 0 fold change (red line) on the left side of the graph is due to a DESeq2 $\log_2$ prior, which moderates estimates from genes with low or highly variable counts. Points that fall outside of the window are plotted as upwards or downwards facing triangles.

5.3.4 Temporal DE

The count data of 287,421 transcripts under LD and LL were analyzed for temporal DE. These datasets were independently filtered by removal of genes with low normalised counts; only genes with a large normalised count contained sufficient information to yield a significant call during dispersion estimation (Figure 2). Temporal DE was only assigned to transcripts with an adjusted p value (padj) ≤ 0.01. p value adjustment was performed using the Benjamini Hochberg method given by padj = p x (n/n–r), where n is the number of tests performed and r is the rank of p when p values from all tests are sorted in ascending order. For example, the second largest p value will be multiplied by n/n-1 and the third largest by n/n-2 (Noble, 2009). The number of transcripts excluded from the DE or non- DE lists due to independent filtering was 229,953 from the LL dataset and 143,806 from the LD dataset. The total number of transcripts that remained in the LD dataset was 143,615 and 57,468 in the LL. Altogether 201,083 transcripts from LL and LD were included in the following outputs of DE analysis.

15,662 (11%) transcripts were temporally differentially expressed in the LD dataset compared to 3,676 (6.39%) in the LL dataset. 3,053 (1.5%) transcripts were temporally differentially expressed in both LL and LD datasets (although this figure is calculated from the number of transcripts differentially expressed in both LD and LL conditions, divided by the additive populations of the LD and LL datasets, of which the LL dataset comprised a far smaller proportion). 10,428 transcripts were differentially expressed in LD conditions but not LL, *i.e.* were photoperiodic. 16,285 transcripts were differentially expressed under one or both conditions of LL and LD.

Of the 201,083 transcripts analyzed in both light regimes, 181,743 (90%) had no significant difference in expression levels (*i.e.* were constitutively expressed) over the diel cycle. In LD, 143,615 transcripts in total were analyzed, and 127,952 (89%) had no significant DE over the time course. 53,791 (93%) transcripts had no significant DE in the LL dataset.

5.3.4.1 Constitutively expressed genes and pathways

Using BLAST annotation data and MEGAN software (section 3.3.3.1), *Viridiplantae* transcripts were extracted from the BLAST annotations of the whole *B. braunii* transcriptome assembly. In the following text, 5-digit KEGG pathway map identifiers are stated in brackets.

A KEGG Global Metabolic Pathway map (01100) of the *Viridiplantae* transcripts that had no detectable temporal DE was generated using Annot8r assigned KEGG Orthologies (KOs) (section 3.2.4.2) and the KEGG Search and Colour online tool (http://www.genome.jp/kegg/tool/map_pathway2.html).

Areas of pathways sparsely mapped in the overview Metabolic pathways map are Metabolism of Cofactors and Vitamins, Biosynthesis of Secondary Metabolites, Amino Acid Metabolism, Xenobiotics and Biodegradation and Metabolism and lastly Metabolism of Terpenoids and Polyketides (Figure 3). Unmapped edges in the whole metabolome map occur where enzymes are present to synthesize the nodes either side but not convert between the two nodes, creating a gap on the map.

642 KOs from non-temporally differentially expressed transcripts were mapped to the KEGG Global Metabolic Pathway map (01100). 236 components involved in Biosynthesis of Secondary Metabolites (01110) were identified that were not differentially expressed.

Sets of pathways lacking coverage by the non-differentially expressed dataset of *B. braunii* transcripts were Glycan Biosynthesis and Metabolism, Xenobiotics Biodegradation and Metabolism, Metabolism of Cofactors and Vitamins, particularly the latter two. Lipid biosynthesis pathways with notable gaps included Steroid biosynthesis (00100) and Steroid hormone biosynthesis (00140).

55 proteins were identified with roles in the Cell Cycle pathway (04110), although this left 35 components unaccounted for. 24 of the 35 KEGG listed Eukaryotic basal transcription factors (03022) were identified with no temporal DE.

The citrate cycle (00020) was incomplete with only 8 components identified, leaving 20 absent. Gaps remained in other group subdivisions of Carbohydrate Metabolism and Energy Metabolism as well. 59 oxidative phosphorylation pathway (00190) components were mapped and 13 protein subunits and enzymes of the Photosynthesis pathway (00195) were identified, although the majority (59) remained unmapped. 25 Fatty Acid Metabolism (01212) enzymes were identified from pathways localised to the cytoplasm or plastid and mitochondria and endoplasmic reticulum. Gaps in Fatty Acid Metabolism pathways were observed mainly in pathways localised

to endoplasmic reticulum. 9 Carbon Fixation in photosynthetic organisms (00710) pathway components were identified as having no temporal DE.

Aside from the overview pathway categories, pathways with the highest number of KOs mapped include purine metabolism (00230) with 94 KOs, Spliceosome (03010) had the second highest number of KOs; 90, followed by Ribosome (03010) with 74. Other highly mapped pathways were RNA transport (03013), within which 70 components were identified and 69 and 68 KOs from Pyrimidine biosynthesis (00240) and Ubiquitin mediated proteolysis (04120) also.

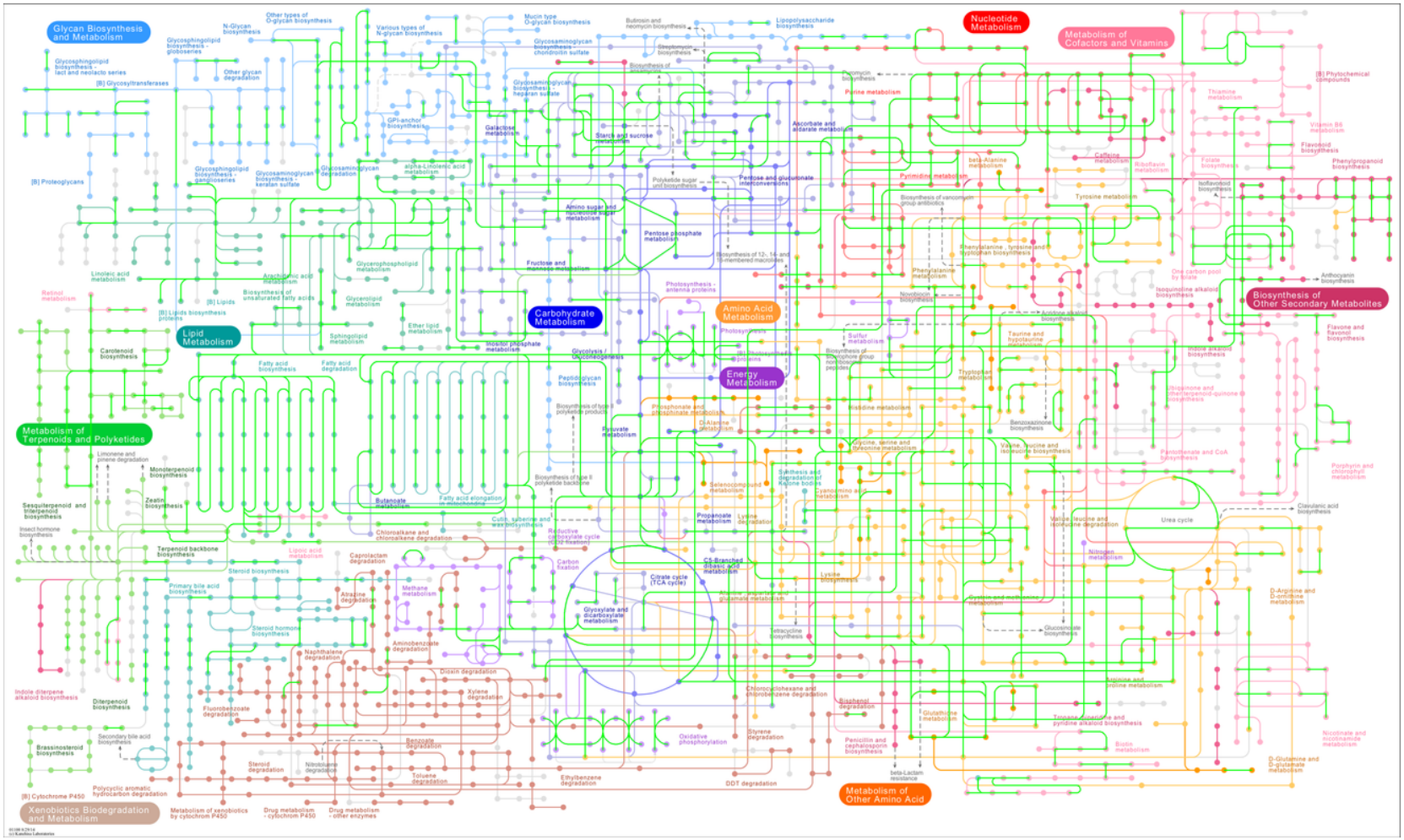


Figure 3 KEGG Metabolic pathways overview map of *B. braunii* constitutively expressed transcripts

Constitutively expressed (padj ≤ 0.01) *B. braunii* trancsripts under LL and LD conditions are shown (green lines) mapped by KO identifier to enzymes and proteins connecting nodes, which represent compounds.

12 KOs were assigned to transcripts that mapped to the terpenoid backbone synthesis pathway (00900) enzymes. Gaps in the 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway (MEP/DOXP pathway) route into squalene synthesis arose where 1-deoxy-D-xylulose-5-phosphate synthase (01662), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (00099), 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [00991), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (01770), (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (03526), isopentenyl-diphosphate delta-isomerase (01823) and isoprene synthase (12742) remained unmapped by non-differentially expressed transcripts (Figure 4). MEP/DOXP enzymes identified in the non-differentially expressed dataset were 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (00919), 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (03527) and farnesyl diphosphate synthase (00787). Squalene synthase (00511) was also mapped using KO assignment to transcripts.

(2Z,6Z)-farnesyl diphosphate synthase (15887) and farnesyl diphosphate synthase (00787) were both missing from the Terpenoid pathway also. However short-chain Z-isoprenyl diphosphate synthase (12503) was present.

However farnesyl-diphosphate farnesyltransferase (00801) was missing from the part of the Sesquiterpenoid and triterpenoid synthesis pathway (00909) responsible for squalene synthesis (Figure 5).

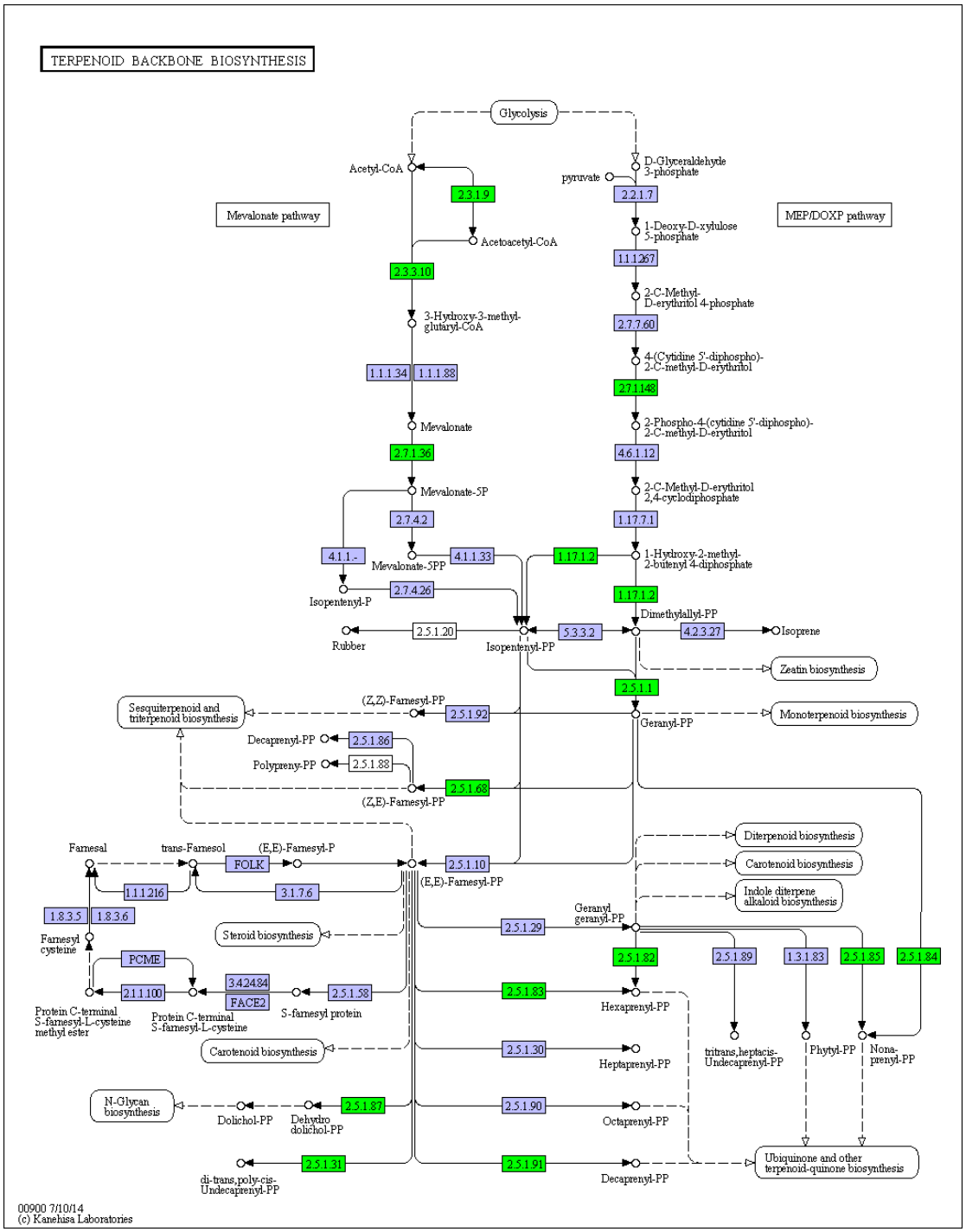


Figure 4 KEGG Terpenoid backbone pathway mapped with *B. braunii* constitutively expressed transcripts
B. braunii constitutively expressed transcripts (padj≥0.01) mapped using KO identifier to the enzymes (green boxes) of the KEGG terpenoid backbone reference pathway. Enzymes unmapped by *B. braunii* constitutively expressed transcripts are shown in blue. Outputs to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

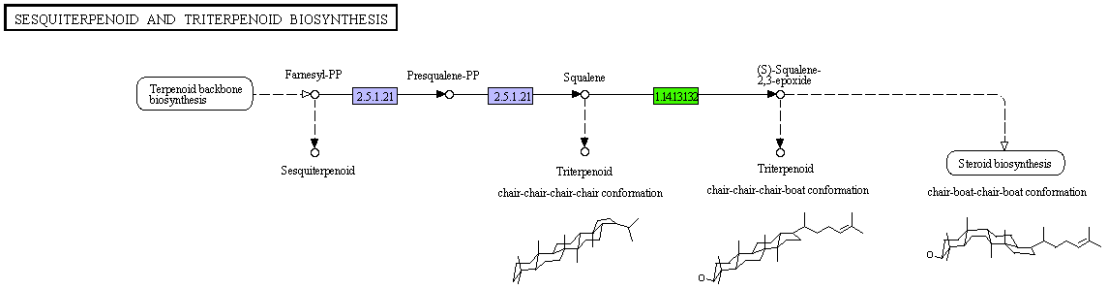


Figure 5 KEGG Sesquiterpenoid reference pathway with constitutively expressed *B. braunii* transcripts mapped

B. braunii constitutively expressed (padj≥0.01) mapped using KO identifier to the enzymes (green boxes) of the KEGG sesquiterpenoid and triterpenoid reference pathway. Enzymes unmapped by *B. braunii* constitutively expressed transcripts are shown in blue. Outputs to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

5.3.4.2 Genes and pathways with DE

3,053 transcripts were assigned temporal DE (padj≤0.01) in both LD and LL. After using BLAST results to filter to only those transcripts within the *Viridiplantae*, 240 differentially expressed transcripts were mapped on to the KEGG Global Metabolic Pathway map (01100) with KO identifiers (Figure 6). 111 of the mapped transcripts were involved in Biosynthesis of Secondary Metabolites (01110).

Areas less covered by temporally differentially expressed transcripts in the Global Metabolic pathways map are Metabolism of Cofactors and Vitamins, Biosynthesis of Other Secondary Metabolites, Amino Acid Metabolism, Xenobiotics and Biodegradation and Metabolism. Sets of pathways that had good coverage by differentially expressed *B. braunii* transcripts were Metabolism of Terpenoids and Polyketides, Lipid Metabolism, Carbohydrate Metabolism and Energy Metabolism.

There were 40 KOs assigned to temporally differentially expressed transcripts that mapped within Carbon Metabolism pathways (01200)- the highest number of KOs assigned to a single group of pathways in this dataset. Protein processing in the endoplasmic reticulum (04141) was well supported by KO annotated transcripts within the DE dataset, with 24 transcripts mapped. Fatty acid metabolism (01212), Purine metabolism (00230), Biosynthesis of amino acids (01230), Starch and sucrose metabolism (00500) and Cell cycle (04110) were amongst the top most supported groups of pathways with between 20 and 24 transcripts assigned to each. Transcripts were assigned with 13 of the Citrate cycle (00020) enzyme KO identifiers, leaving 15 missing from the pathway.

No transcription factors were identified by KEGG mapping in the temporally differentially expressed dataset.

15 of 29 Carbon fixation in photosynthetic organisms pathway (00710) components were identified in the temporally differentially expressed dataset . Just 13 protein subunits and enzymes of the Photosynthesis pathway (00195) were identified, leaving a further 50 unidentified. 23 transcripts were annotated with KOs allocated within Fatty Acid Metabolism (01212), identifying components of the involved pathways localised to the cytoplasm and plastids, mitochondria and to a lesser extent the endoplasmic reticulum. Gaps remaining in Lipid Metabolism pathways include the many enzymes missing from the Steroid biosynthesis pathway, with only seven transcripts mapped with KOs within this pathway.

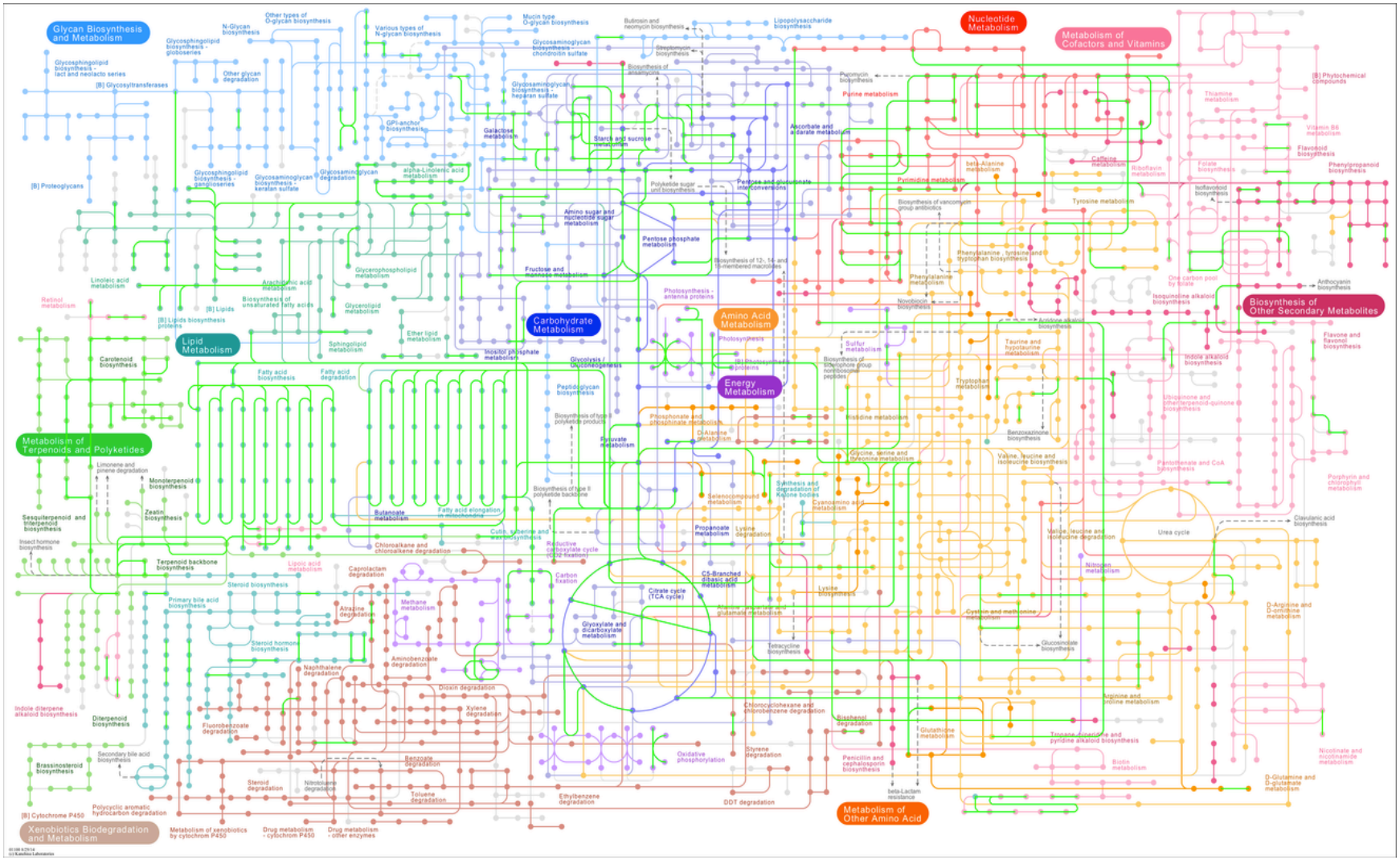


Figure 6 KEGG Metabolic pathways overview mapped with differentially expressed *B. braunii* transcipts
B. braunii transcripts that were differentially expressed (padj≤0.01) under both LD and LL conditions are shown (green lines) mapped by KO identifier to enzmes and proteins connecting nodes, which represent compounds.

From the KOs assigned to temporally differentially expressed *B. braunii* transcripts, 11 terpenoid backbone synthesis pathway (00900) enzymes were identified- all belonging to the 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate pathway (MEP/DOXP pathway) route into squalene synthesis (Figure 7). MEP/DOXP enzymes identified in the dataset were 1-deoxy-D-xylulose-5-phosphate synthase (01662), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (00099), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (00919), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (01770), (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (03526) and 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (03527). 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (00991) and isopentenyl-diphosphate delta-isomerase (01823) were missing from the MEP/DOXP pathway, as was isoprene synthase (12742).

Amongst the temporally differentially expressed *B. braunii* dataset were 3 KO identifiers that mapped to squalene monooxygenase (14.13.132), farnesyl-diphosphate farnesyltransferase (2.5.1.21) and epi-isozizaene 5-monooxygenase (1.14.13.106) of the Sesquiterpenoid and triterpeniod pathway (Figure 8). 16 *B. braunii* transcripts were annotated with a KO identifier of one of the aforementioned three enzymes. *B. braunii* comp174730_c0_seq1, comp175398_c0_seq1, comp175446_c0_seq1 and comp176380_c0_seq1 were annotated as squalene synthase with E values of 5.00e-32, 6.00e-82, 1.00e-90 and 1.00e-119 respectively. Sequences 1, 3, 5 and 7 of comp174730_c0 from the temporally differentially expressed data were all assigned to squalene synthase by KO identifier.

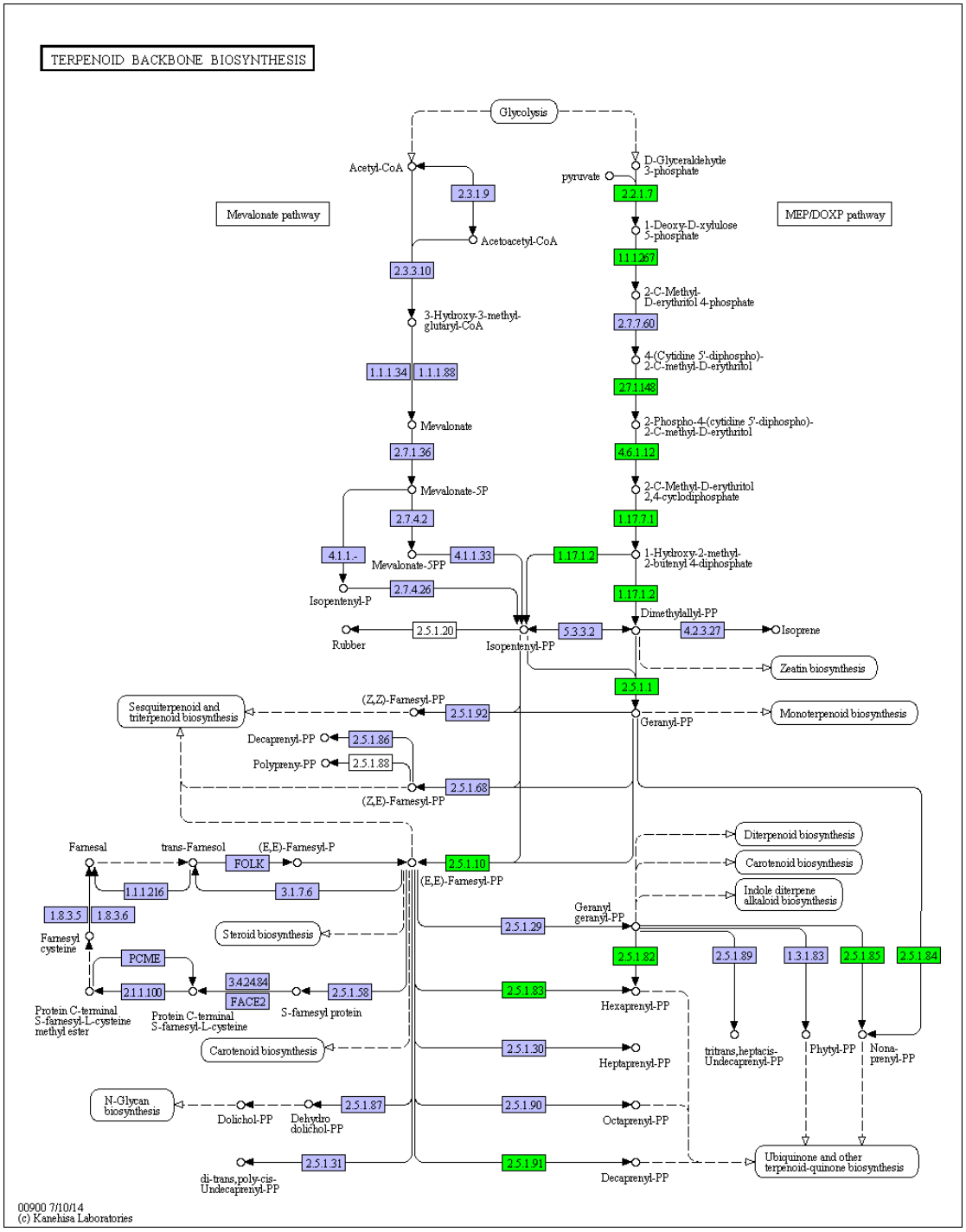


Figure 7 KEGG Terpenoid backbone synthesis pathway with differentially expressed *B. braunii* transcripts
B. braunii transcripts that were differentially expressed under both LL and LD (padj ≤ 0.01) mapped with KO identifier to the enzymes (green boxes) of the KEGG terpenoid backbone reference pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

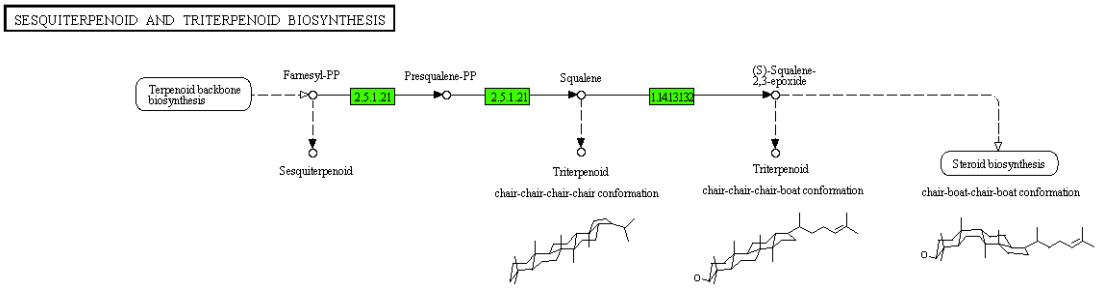


Figure 8 KEGG Sesquiterpenoid pathway mapped with differentially expressed *B. braunii* transcripts

B. braunii transcripts that differentially expressed under both LD and LL conditions (padj ≤ 0.01) mapped with KO identifier to the enzymes of the KEGG sesquiterpenoid reference pathway. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

9 components of the KEGG Circadian Clock (plant) pathway (04712) were identified in the set of *B. braunii* transcripts that were temporally differentially expressed under both LD and LL conditions. Identified components were CCA1 (12134), LHY (12133), PRR7 (12129), ZTL (12115), PRR3 (12131), FKF1 (12116), CO (12135), CHS (00660) and CK2 alpha chain (03097). TOC1, PRR5, PRR9, ELF3 were amongst those that remained unidentified within the differentially expressed dataset.

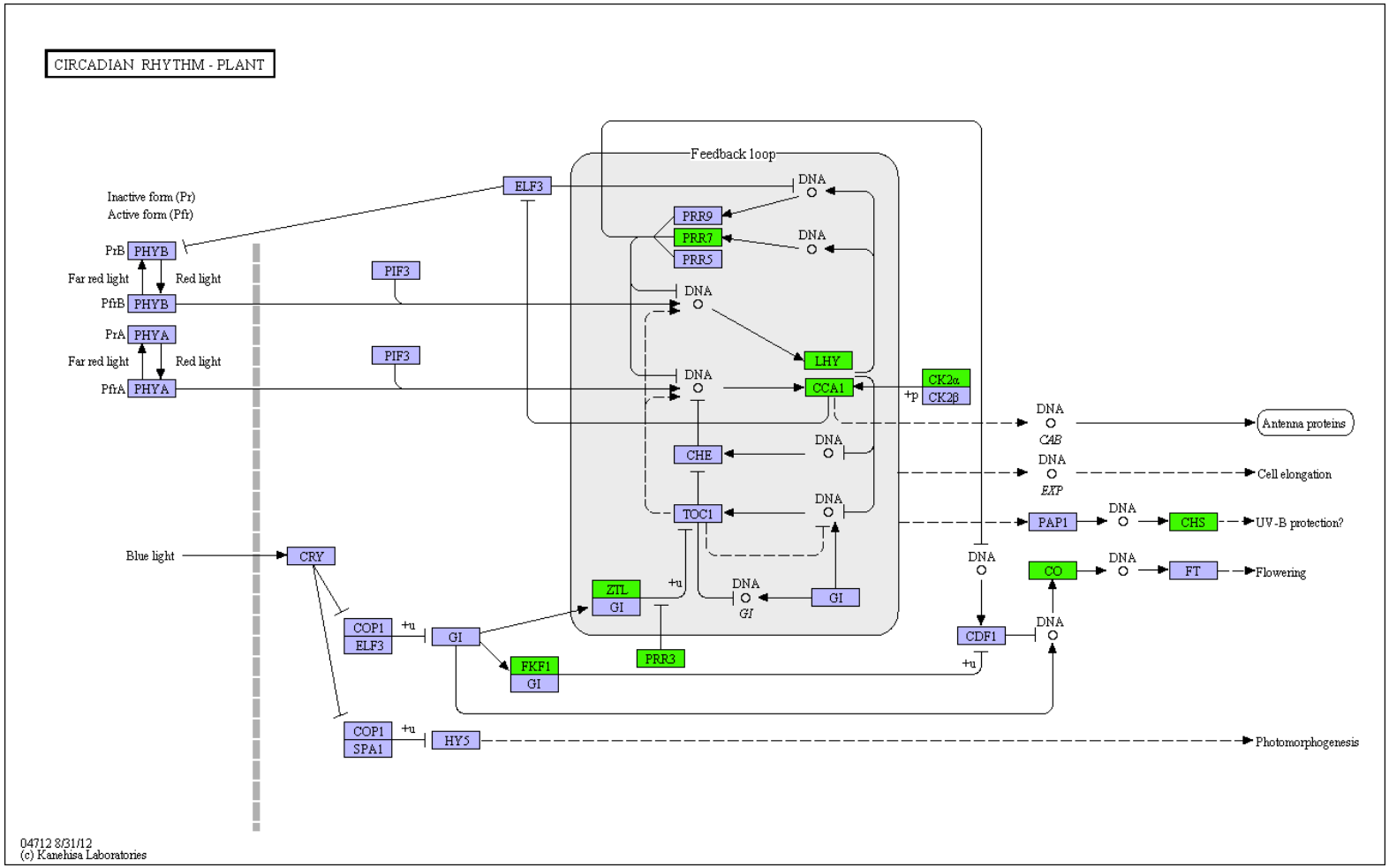


Figure 9 KEGG Circadian clock (plant) pathway mapped with *B. braunii differentially* expressed transcripts

B. braunii transcripts with differential expression (padj ≤ 0.01) in both LD and LL conditions, mapped with KO identifier to the enzymes (green boxes) of the KEGG Circadian rhythm (plant) reference pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes.

5.3.4.3 Genes with light dependent differential expression

In a comparison of the 15,662 differentially expressed transcripts in LD and 53,791 non-DE transcripts in LL (padj≤0.01), 10, 428 transcripts were present in both lists. Of the 10, 428 transcripts that had significant DE in LD but not LL conditions, 599 assigned KO identifiers could be mapped to the KEGG Metabolic Pathway overview map (Figure 10). 242 proteins and enzymes involved in Biosynthesis of Secondary Metabolites (01110) were identified.

Groups of pathways that were sparsely populated by mapped KOs in the photoperiodic dataset were Metabolism of Cofactors and Vitamins, Biosynthesis of Other Secondary Metabolites, Xenobiotics and Biodegradation and Metabolism and Glycan Biosynthesis and Metabolism. The citrate cycle remained incomplete; missing components included aconitate hydratase (01681), citrate synthase (01647), fumarate hydratase class I (01676), ATP citrate (pro-S)-lyase (01648), malate dehydrogenase (quinone) (00116), succinyl-CoA:acetate CoA-transferase (18118), 2-oxoglutarate ferredoxin oxidoreductase subunit alpha (00174) and isocitrate-homoisocitrate dehydrogenase (17753).

Sets of pathways that had good coverage by KO assigned to photoperiodic *B. braunii* transcripts were Metabolism of Terpenoids and Polyketides, Lipid Metabolism, Carbohydrate Metabolism and Energy Metabolism.

81 components of Carbon Metabolism pathways (01200) were identified- the second highest number of KOs assigned to a single group of pathways in this dataset. in photosynthetic organisms components were identified. The highest number of KOs assigned to a reference pathway was 96, to the Ribosome pathway (03010). Other highly annotated pathways were Biosynthesis of amino acids (01230), Protein processing in endoplasmic reticulum (04141), Spiceosome (03040) Purine metabolism (00230) and Pyrimidine metabolism (00240) all with between 39 and 80 KOs mapping to them.

5 of the KEGG listed Basal transcription factors (Eukaryotes) were identified in the photoperiodic dataset.

The photoperiodic dataset contained transcripts assigned with 15 KOs of 29 that are involved in the Carbon fixation in photosynthetic organisms pathway (00710). 17 protein subunits and enzymes of the Photosynthesis pathway (00195) were identified, leaving a further 48 unidentified. 18 transcripts were annotated with KOs allocated within Fatty Acid Metabolism (01212), identifying components localised to the cytoplasm and plastids, mitochondria and endoplasmic reticulum although numerous gaps were observed in the mitochondrial and endoplasmic reticulum compartments. 11

components of the Steroid biosynthesis pathway (00100) were missing, with 18 KOs mapped.

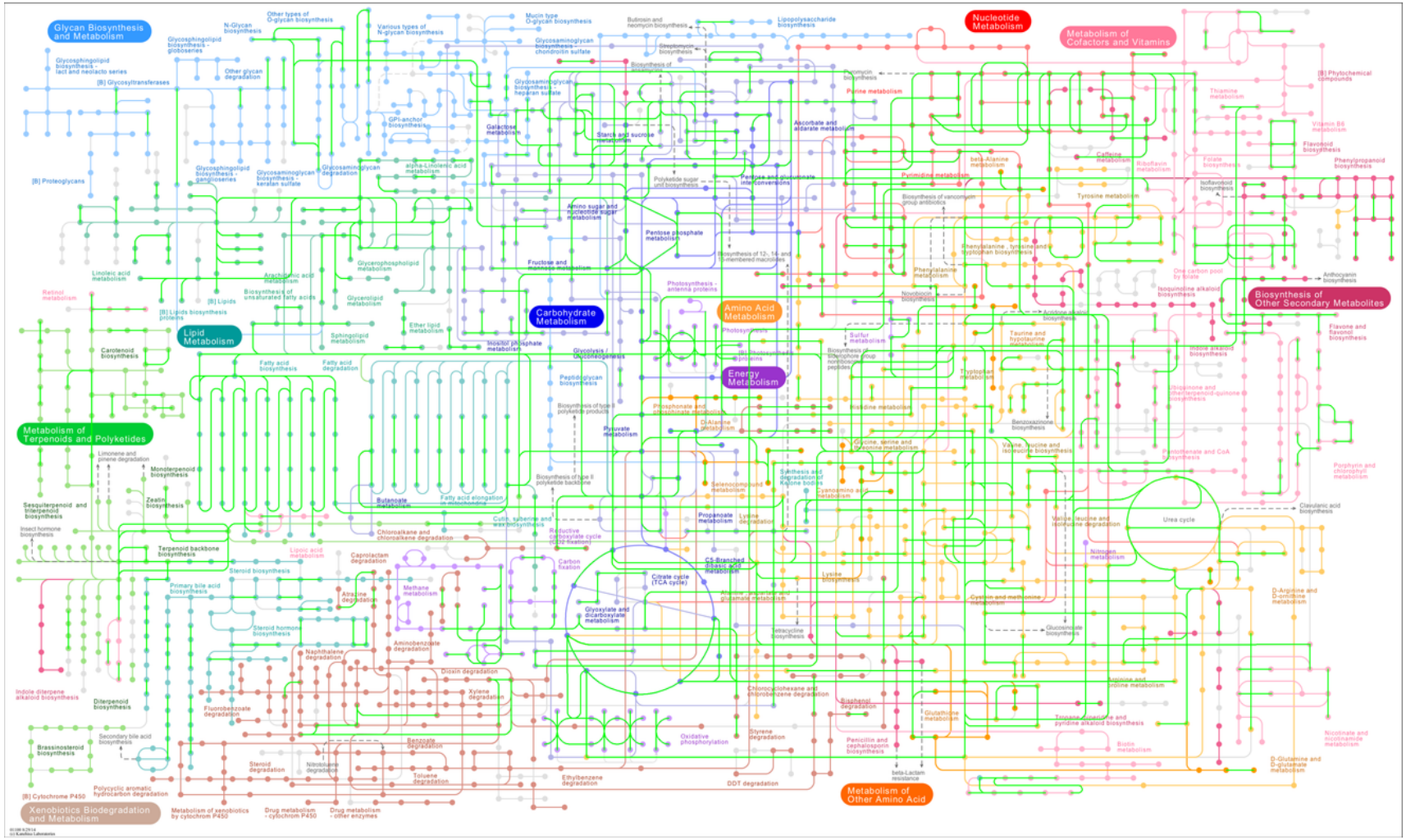


Figure 10 KEGG Metabolic pathways overview with photoperiodic *B. braunii* transcripts mapped
B. braunii transcripts with differential expression under LD conditions but not LL (padj ≤ 0.01) are shown (green lines) mapped by KO identifier to enzymes and proteins connecting nodes, which represent compounds.

Transcripts involved in the terpenoid backbone synthesis pathway were also present in the light dependent DE dataset (photoperiodic), with 6 components identified (Figure 11). Transcripts were mapped to 1-deoxy-D-xylulose-5-phosphate synthase (01662), 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (00991), 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (03527), isopentenyl-diphosphate delta-isomerase (01823) and farnesyl diphosphate synthase (00787) which collectively form part of the MEP/DOXP pathway, although some unmapped genes were required to complete the chain. Missing components of the MEP/DOXP pathway were 1-deoxy-D-xylulose-5-phosphate reductoisomerase (00099), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (00919), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (01770), (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (03526) (2Z,6Z)-farnesyl diphosphate synthase (15887)and short-chain Z-isoprenyl diphosphate synthase (12503). No components of the Mevalonate pathway of the Terpenoid Backbone synthesis pathway were identified.

Photoperiodic transcripts were annotated with KOs associated with farnesyl-diphosphate farnesyltransferase (EC:2.5.1.21), also known as squalene synthase. From the photoperiodic dataset, transcript comp174730_c0_seqs (3, 5, 6, 7, 8) were all annotated by KO as squalene synthase with BLAST E values of $1e^{-44}$, $3e^{-32}$, $2e^{-64}$, $4e^{-24}$ and $2e^{-36}$ respectively (Figure 12).

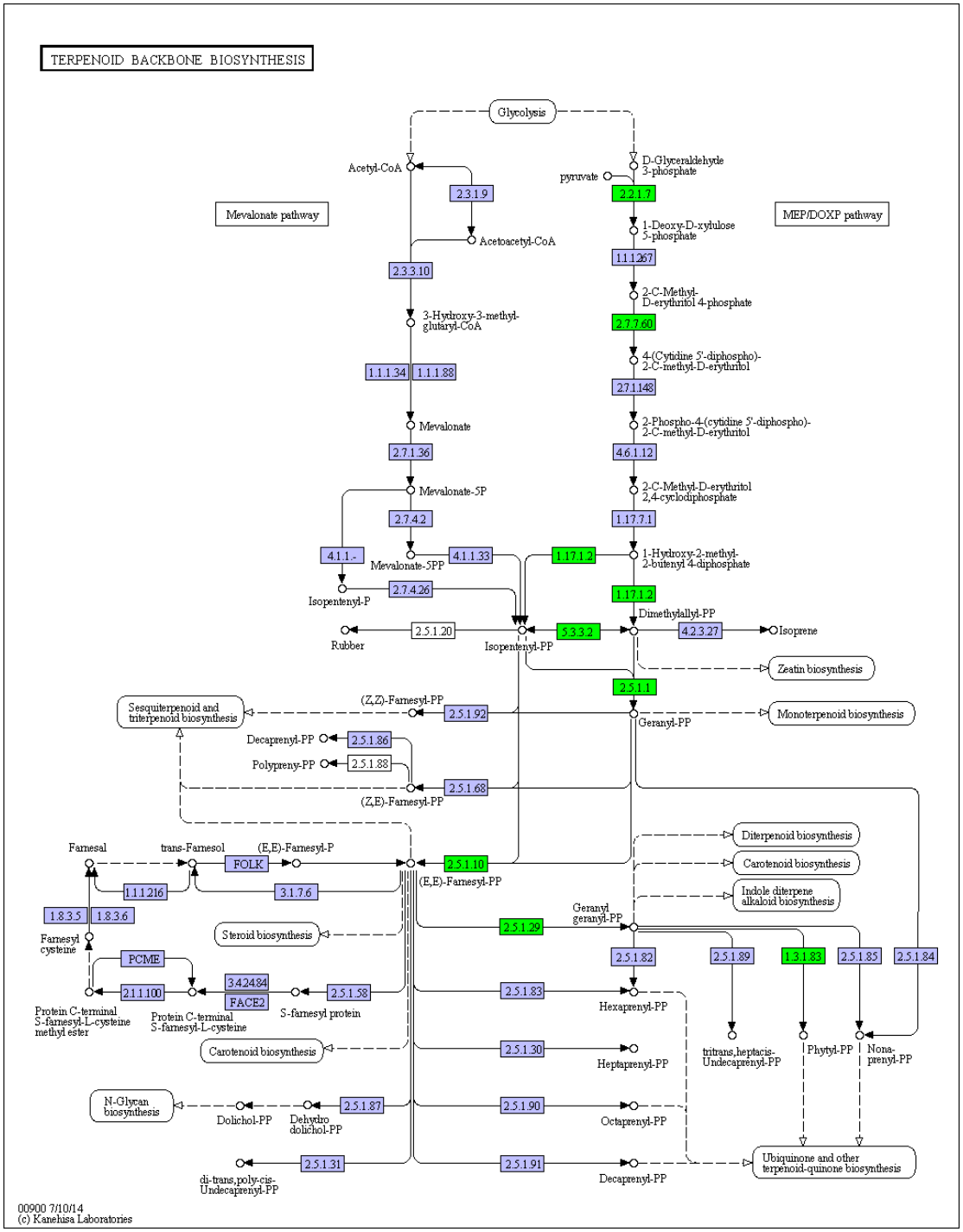


Figure 11 KEGG Terpenoid backbone biosynthesis pathway mapped with photoperiodic *B. braunii* transcripts

B. braunii transcripts with temporal DE (padj ≤ 0.01) in only LD conditions, mapped with KO identifier to the enzymes (green boxes) of the KEGG terpenoid backbone reference pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

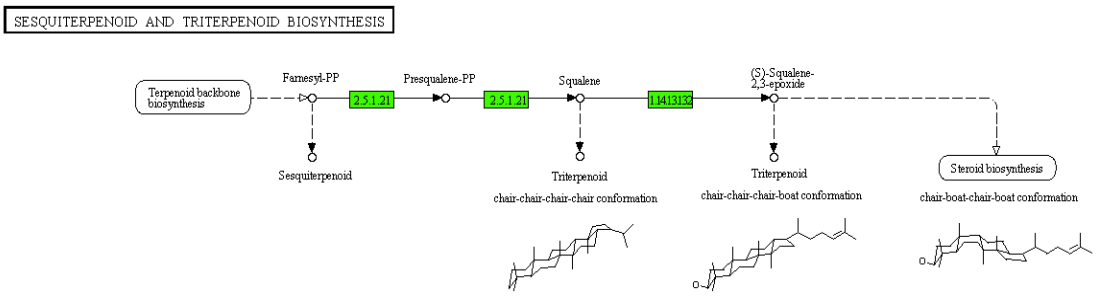


Figure 12 KEGG Sesquiterpenoid pathway mapped with photoperiod *B. braunii* transcripts
B. braunii transcripts with temporal DE (padj ≤ 0.01) in only LD conditions mapped with KO identifier to the enzymes (green boxes) of the KEGG Sesquiterpenoid and triterpenoid reference pathway. Outputs to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

11 *B. braunii* gene products of the KEGG plant circadian clock reference pathway (04712) were identified in amongst *B. braunii* transcripts that were differentially expressed under LD conditions but not LL (Figure 13). LHY (12133) and CCA1 (12134), PRRs 3 (12131), 5 (12130), 7 (12129), 9 (12128) and TOC1 (12127) were all identified alongside other components in the periphery of the clock mechanism including CHS (00660), COP1 (10143), ZTL (12115) and FKF1 (12116).

Sequences 9 and 12 of comp170985 were mapped to LHY (E values 4.00E-19 and 5.00E-19) and CCA1 (both E values of 3.00E-18). Only two transcripts within the photoperiodic dataset mapped to TOC1; comp56012_c0_seq1 and comp56012_c0_seq2 (*BbPRR*), both with an E value of $3.00e^{-21}$.

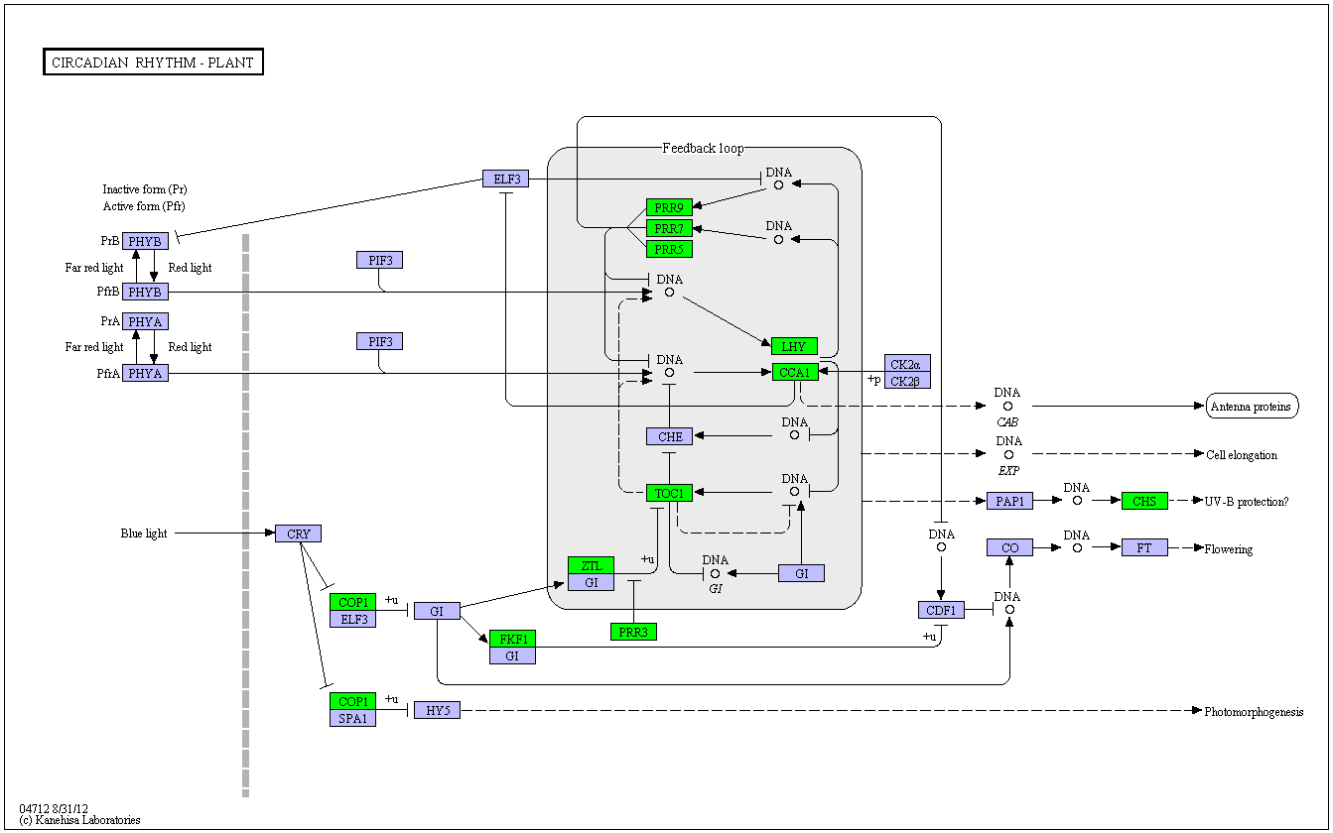


Figure 13 KEGG Circadian clock (plant) pathway mapped with photoperiodic *B. braunii* transcripts
B. braunii transcripts with temporal DE (padj ≤ 0.01) in only LD conditions, mapped with KO identifier to the enzymes (green boxes) of the KEGG Circadian rhythm (plant) reference pathway. Enzymes un-identified in the *B. braunii* transcriptome are shown in blue. Outputs of terpenoid synthesis to other pathways are indicated in rounded-edge boxes. Circles indicate compounds.

5.3.5 Expression of predicted Botryococcus braunii *clock components*

Predicted *B. braunii* core clock components *BbCCA1* and *BbPRR* were searched for within both the differentially expressed and non-differentially expressed datasets. Transcript *BbCCA1* had significant temporal DE in both LL and LD conditions (padj≤0.01). Predicted PRR- like clock component, *BbPRR* had significant temporal DE in only LD conditions (padj≤0.01).

5.3.5.1 Expression profile of predicted Botryococcus braunii *PRR1- like transcript*

Normalised counts of *B. braunii* transcript *BbPRR* were extracted from the DESeq2 output for each timepoint under LL and LD conditions and plotted as a mean of the timepoint replicates. Under LD conditions, mRNA abundance increased from 84 counts to a peak of 568, 8 hours after dawn. Counts fell to 170 as the lights went out 12 hours into the timecourse and remained at approximately this level until falling to 78, 20 hours from dawn. The lowest mean counts were observed 24 hours into the timecourse- as the lights came on for a second time. Subsequently transcript abundance increased to 268 counts in the final timepoint, 28 hours from first dawn. The lowest expression value of *BbPRR* was observed at time 0 and the highest 8 hours afterwards.

Under LL conditions, the expression profile of *BbPRR* was somewhat dampened although the shape resembled the diurnal profile. LL counts ranged from 304 to 414 compared to 47 to 568 counts in the LD conditions. The overall counts were up-regulated compared to those under LD conditions, despite the range in expression being smaller. In contrast to LD conditions, the highest abundance of *BbPRR* counts was seen at 24 hours after dawn. The lowest counts of 304 were observed at time 0, following this there was a small increase to 377 after which counts fell moderately to 325 at 8 hours. Counts increased to 335 at 12 hours and then changed little, falling to 314 and up again to 353 at timepoints 16 and 20 hours respectively. Abundance peaked at 414 counts at 24 before falling to 331 in the final timepoint, 28 hours from first “subjective” dawn.

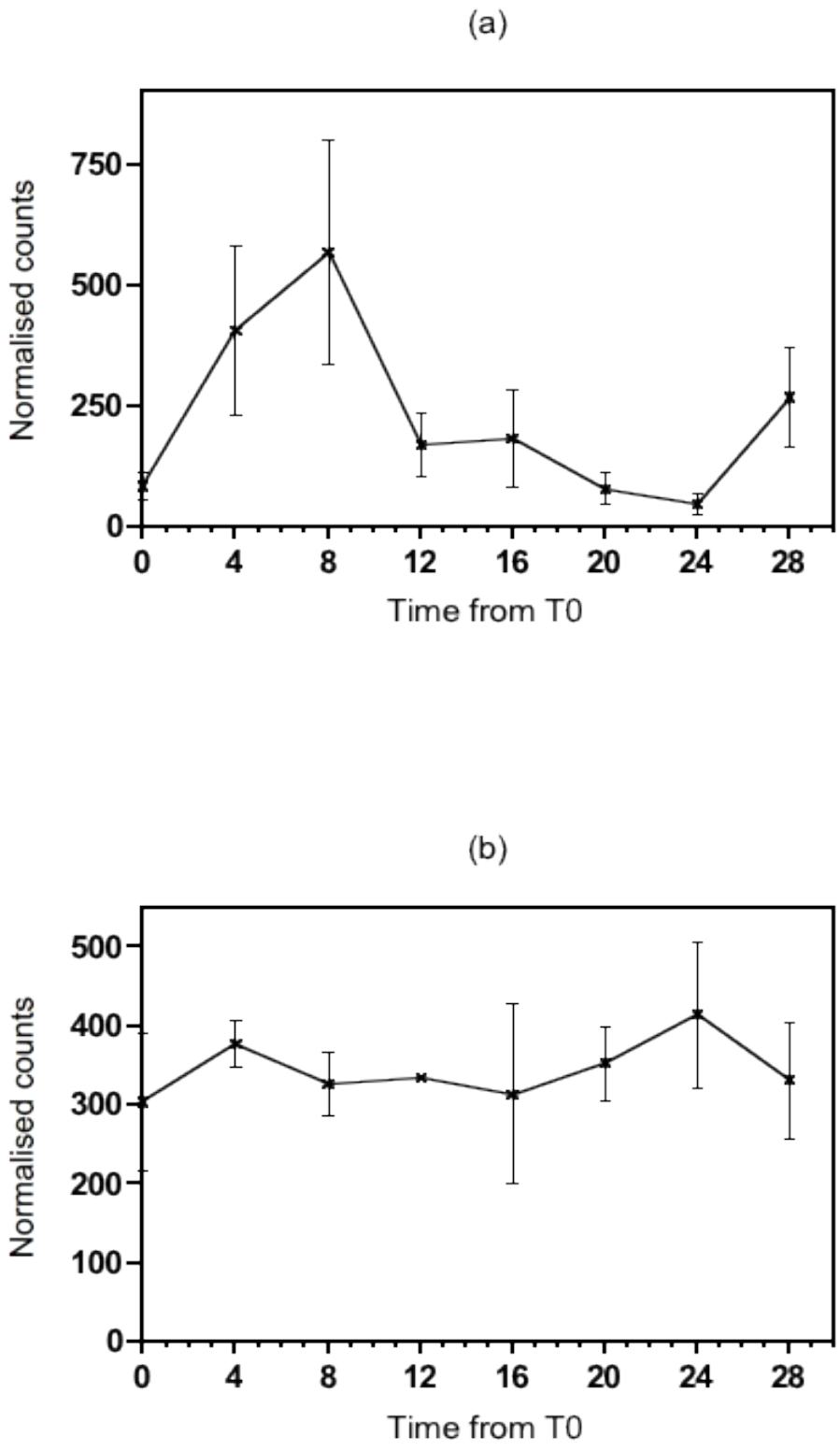


Figure 14 mRNA expression of *BbPRR*
Normalised read counts for transcript *BbPRR* in LD (a) and LL (b). Error bars are plotted from standard deviation values.

5.3.5.2 Expression of predicted Botryococcus braunii CCA1-like transcript

Under LD conditions, mean counts for *B. braunii* transcripts *BbCCA1* are low at the beginning of the time course, with 6 counts, and remain more or less level (4 counts at timepoint 4) until increasing to 35 8 hours after dawn and peaking at 61 counts after 12 hours. Subsequently counts to descend to 28 by the 16th hour from dawn and continue to fall to 10, and subsequently 8 and 3 after 20, 24 and 28 hours respectively. The timepoint with the lowest abundance (4 counts) was 28 hours and the highest abundance (61 counts) was observed at 12 hours from time 0.

The mean abundance of *BbCCA1* commences the time course at 6 counts under LL conditions, increasing to 15 counts after 4 hours. mRNA expression of *BbCCA1* builds to a peak from 41 to 46 counts at timepoints 8 and 12 hours before decreasing to 17 and subsequently 6 counts at timepoint 16 and 20 hours. Abundance begins to rise again at 24 hours (9 counts) and 36 counts are observed in the final timepoint of 28 hours.

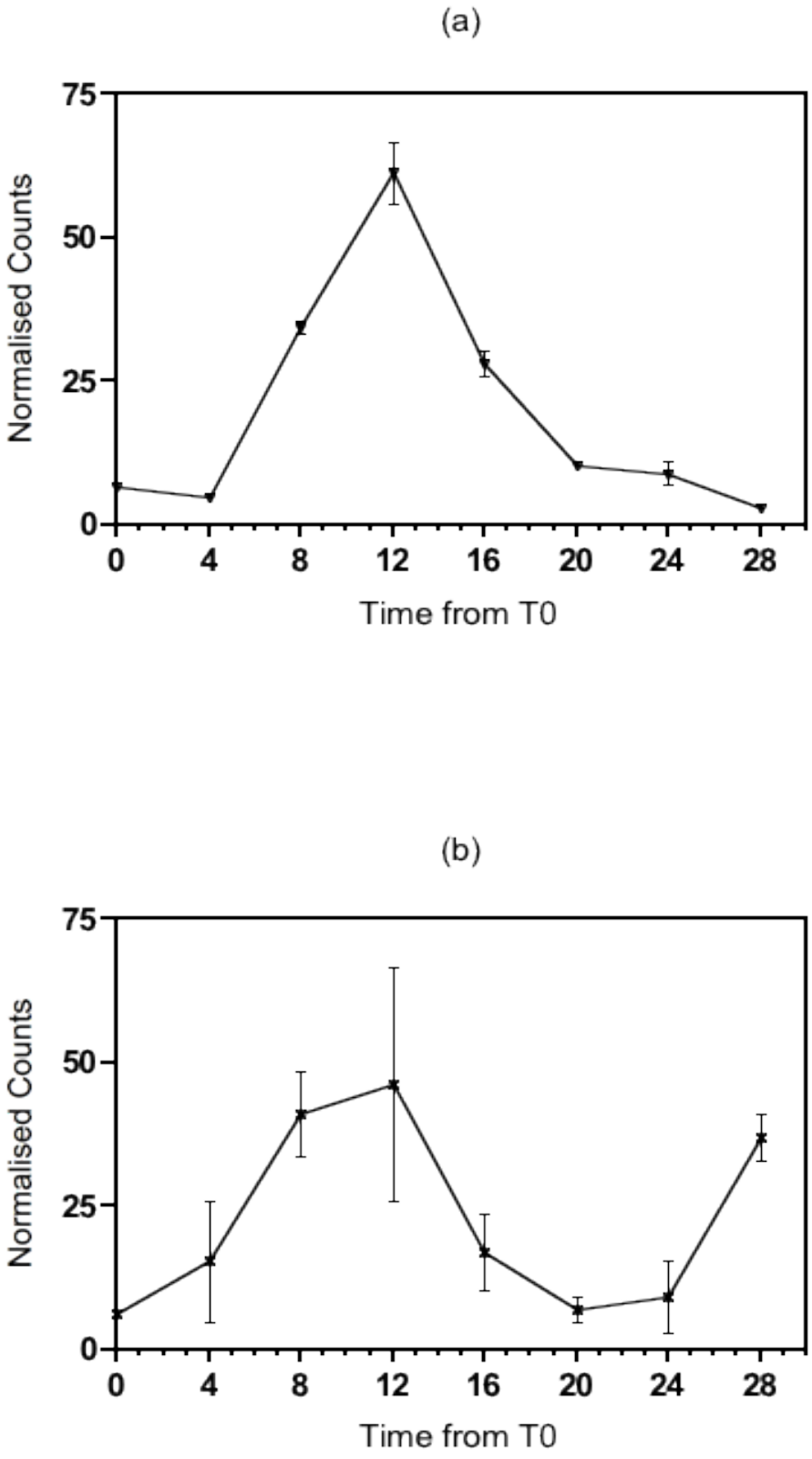


Figure 15 mRNA expression of *BbCCA1*
mRNA expression of *BbCCA1* in LD (a) and LL (b). Error bars are plotted from standard deviation values.

5.3.5.3 Comparative analysis of expression patterns

mRNA expression data for model clock genes collected under 12/12 photoperiods and/ or constant light depending on availability were mined from the literature. Relative expression pattern profiles were generated for each gene by transforming raw values into a percentage of individual maximum expression. The *B. braunii* raw data was processed in the same way and plots of relative expression were created manually. *B. braunii* timepoint intervals were 4 hours, whereas data taken from studies of other organisms were at intervals of 3 hours. The *B. braunii* dataset is 28 hours long, incorporating 2 timepoints, that overlap if considering a 24 hours cycle; samples taken at +24 and +28 hours from dawn overlap with those taken at 0 and +4 hours from dawn.

Under the LD conditions the overall expression pattern of *B. braunii BbPRR* was similar to that of the other *TOC1* mRNA profiles although was the earliest to peak at 8 hours after theoretical dawn, followed by *A. thaliana*, *O. tauri* and *O. sativa TOC1* transcripts 9 hours after dawn (Figure 16 (a)). The amplitude of the *B. braunii* oscillation was approximately 90% deviation from the maximum expression level. The *BbPRR* wave of expression was wider at the base than others, taking a full 24 hours to peak and trough. The ascent of *B. braunii* expression to its peak was more gradual than other profiles, descent was at a similar rate until 12 hours after dawn when the rate became slower creating a slight shoulder in the wave of expression.

O. tauri TOC1 expression had a distinctive shape; ascending rapidly 4 hours after dawn to a wide peak and falling into a rapid descent 12 hours after dawn. Also uniquely, the *A. thaliana* expression pattern had a smaller secondary peak 21 hours after dawn, which was not observed in *B. braunii* nor the other *TOC1* profiles.

In conditions of constant light *BBPRR* expression pattern oscillation was dampened, deviating a maximum of 30% (Figure 16 (b)). Under constant light, *BbPRR* exhibited little phase shift within the first 12 hours, compared to expression under LD conditions. However, expression levels begin to rise again at 16 hours after dawn to a peak at 24 hours after dawn representing a forwards phase shift of 8 hours in comparison to expression under LD conditions.

Contrastingly, *A. thaliana* and *O. sativa TOC1* profiles maintained amplitude near to that under LD conditions. *O. tauri TOC1* expression data was not available from constant light conditions.

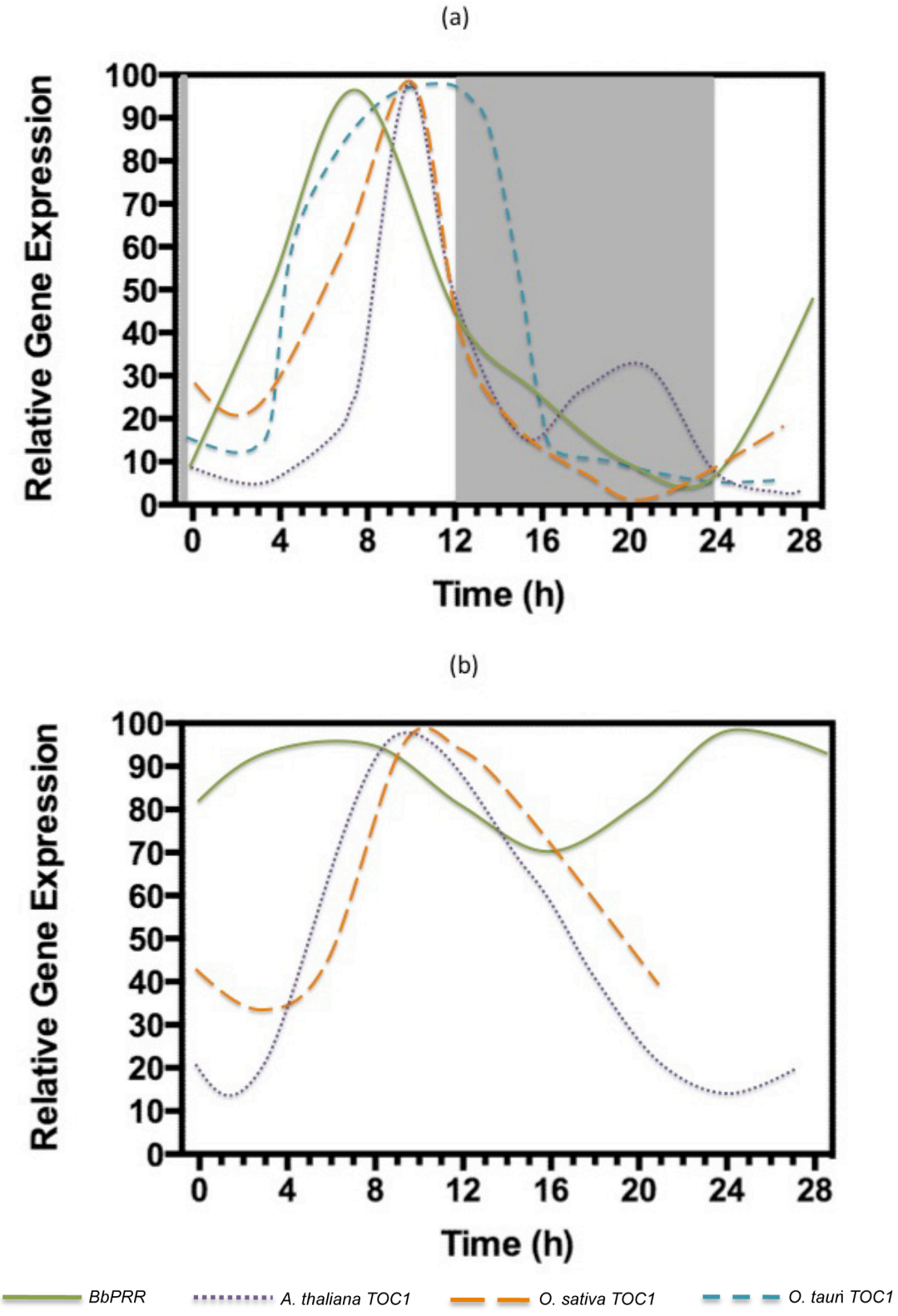


Figure 16 Expression patterns of *TOC1*

Relative mRNA expression patterns of *BbPRR* and it's *TOC1* homologs in LD (a) and LL (b). Expression data is transformed to percentages of the maximum expression for each gene/ transcript.

Under LD conditions, the abundance of *BbCCA1* mRNA remains low from time 0 to 4 hours, after which it climbs to a peak at 12 hours before falling again to approximately 10% of maximum abundance at 20 hours after dawn (Figure 17 (a)). The profile shape of *BbCCA1* is very similar to that of *A. thaliana* and *O. sativa CCA1* but the peak at timepoint 12 is approximately 12 hours earlier in phase, as the latter two peak at 21 and 24 hours from dawn. *O. tauri CCA1* abundance begins to increase after 9 hours from dawn, reaching the peak at 16 hours. The *O. tauri* profile shape is dissimilar to those of *B. braunii*, *A. thaliana* and *O. sativa* with a much broader peak sustained between the 16 hours and 20 hours timepoints. Notably, *O. tauri CCA1* never falls below approximately 30% of the maximum abundance value, whereas *CCA1* from *A. thaliana*, *O. sativa* and *BbCCA1* all fall to approximately 0%, 4% and 10% respectively.

In constant light, BbCCA1 maintains the same phase as in LD conditions although the profile is damped, markedly more so than those and of *A. thaliana* and *O. sativa CCA1* (Figure 17 (b)). An earlier phase shift of approximately 6 hours is observed in the expression of *A. thaliana CCA1.* No data were available for *O. tauri CCA1* mRNA abundance under constant light conditions.

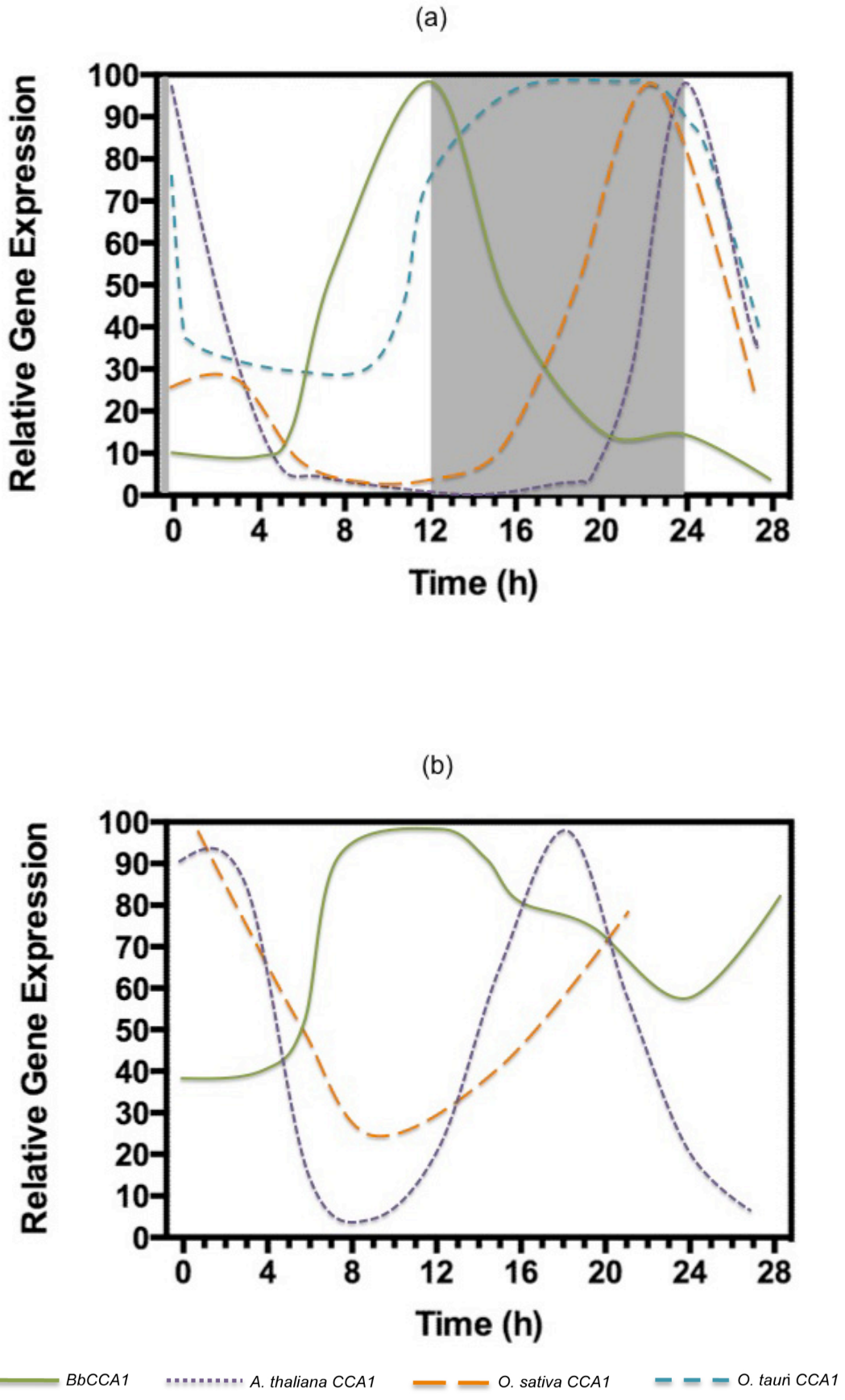


Figure 17 Expression patterns of *CCA1*

Relative mRNA expression patterns *BbCCA1* and it's *CCA1* homologs in LD (a) and LL (b). Expression data is transformed to percentages of the maximum expression for each gene/ transcripts.

5.3.6 Co-expression networks and cluster analysis

5.3.6.1 K-means clustering

Mean expression profiles of transcripts under LD and LL conditions, with padj≤0.001 were clustered into four groups, which were representative of the majority of variance in the expression pattern data.

In *K*-means analysis, *K* (4 in this case) random datapoints were selected as "centroids", from the mean counts at each timepoint. The mean counts were subsequently randomly distributed between the 4 centroids at each timepoint. The point-to-centroid distance was then minimised in an iterative process of the K-means algorithm rearranging datapoints to the centroid with the nearest mean. (Huang, 1998). In the following text, reference to transcript expression is relative to the pattern generated within the respective cluster, as opposed to absolute quantities.

In cluster (a) the expression profiles of transcripts under LD conditions (Figure 18) was high at dawn but fell to much lower abundance 4 hours later. Counts then rose gradually to a peak 16 hours after dawn and fell again slowly to very low levels by 28 hours after dawn. The waveform generated by expression pattern in cluster (a) was broad but of high amplitude. Abundance at dawn was close to that of the 24LD timepoint and abundance at +4 hours from dawn was very close to that at 28 hours.

The waveform of transcripts in cluster (b) had a lower amplitude and was narrower than that of those in cluster (a), with transcripts rising in expression from 4 hours after dawn to a peak at 8 hours and remaining at near peak expression for approximately 4 hours. Between 12 hours after dawn and 16 hours after dawn, the expression of cluster (b) transcripts fell quickly but thereafter continued to fall slowly to reach their lowest levels at 24 hours after dawn. Expression began to increase again at the +28 hours timepoint. Abundance was similar at 0LD and 24LD and 4LD and 28LD.

Cluster (c) transcripts had a high level of expression at dawn and fell with a gradual gradient from 0 to +8 hours and remained at a constant level between 8 and 12 hours. Expression then increased slightly between 12 and 16 hours before increasing more rapidly from 16 to a peak at 24 hours. The peak of expression was sharp, dropping quickly in the following 4 hours to a level similar to that at 4 hours. Expression of cluster (c) transcripts level was a similar level at 0 and 24 hours.

In (d) transcripts clustered together that peaked in expression 4 hours after dawn but only briefly, before falling at a near constant rate, although slightly more rapidly between 4 and 8 hours, until 16 hours after dawn. Abundance gradually increased from 16 hours to 24, at which point a rapid rise in abundance was observed.

There were four predominant patterns of expression within the data. Clusters (a) and (c) had elevated abundance at dawn although cluster (a) peaked at 16 hours from dawn whereas cluster (c) peaked at 24 hours from dawn. The remaining two clusters peaked during the day, one four hours later than the other.

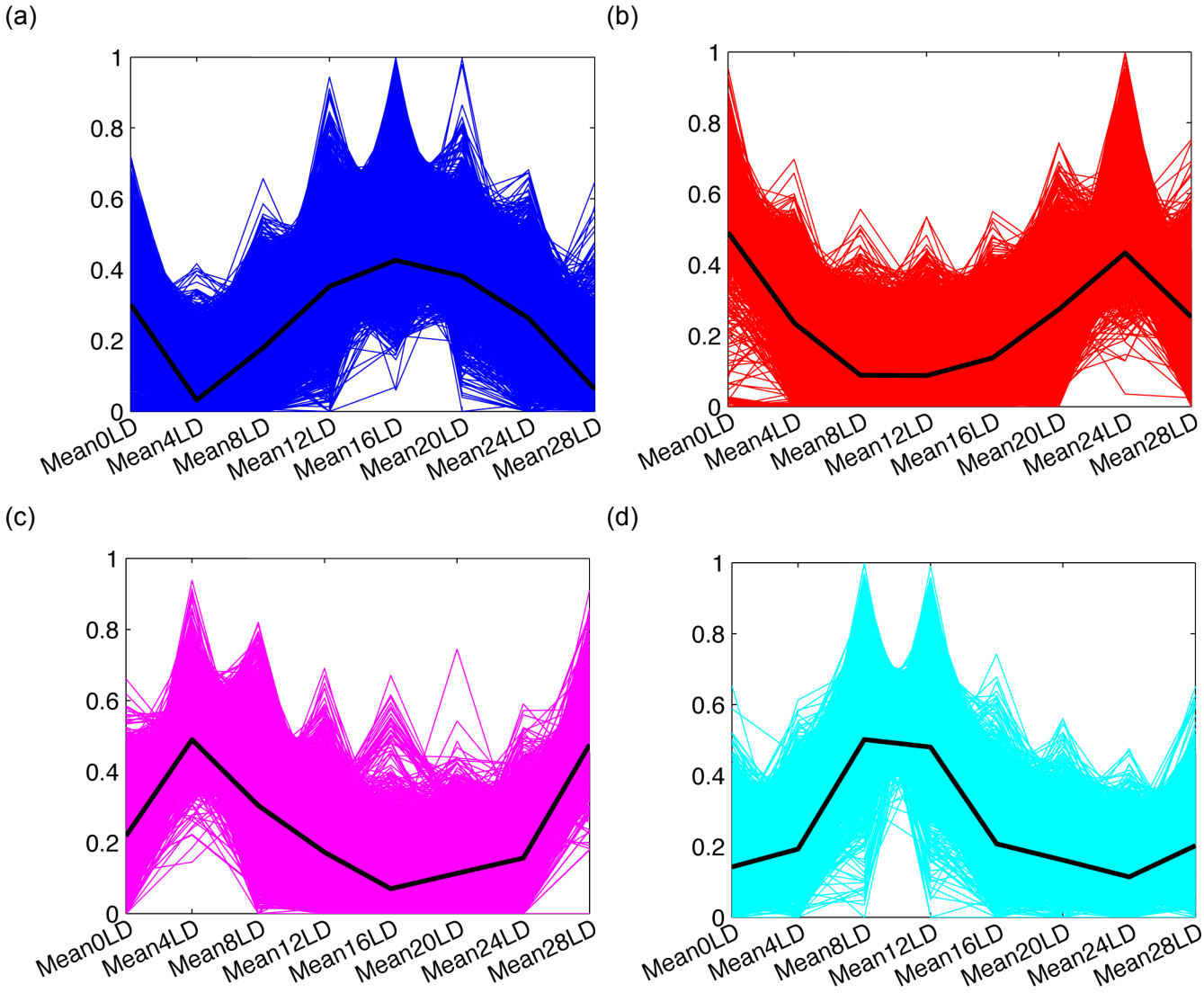


Figure 18 *K*-means clustering of transcripts under LD conditions
The normalised counts of transcripts under LD conditions clustered together according to expression profile similarity by association with the nearest mean. Data are standardised to fit an arbitrary scale on the y axis. The trend of each pattern is represented by a black line.

Transcripts in cluster (a) under LL conditions (Figure 19) were initially low in abundance and rose throughout the morning from 0 to +4 hours when a sharp peak was observed before abundance fell at a near constant rate until 16 hours. Abundance increased again between the hours of 16 and 24, when the profile flattened out for 4 hours until the final timepoint at +28 hours. Expression was elevated in transcripts at time 24 compared to time 0. Peak abundance was observed 28 hours after the first theoretical dawn (time 0), where it was similar to that after only 4 hours from theoretical dawn.

Transcripts clustered together in (b) that had low level expression at time 0 but subsequently increased to a peak 8 hours later. Abundance remained relatively high with a slight decrease before falling from 12 to 20 hours from theoretical dawn. A slight decrease in expression was observed in the pattern between timepoints +20 and 24 hours after which, abundance was elevated again in timepoint 28 hours. The lowest abundance was observed at the first theoretical dawn in cluster (b), subsequent expression levels did not fall as low in timepoint 24 nor any other. Expression levels in timepoint 28, were higher than those in timepoint 4.

In (c) transcript abundance was at its highest level at time 0, thereafter decreasing quickly until 8 hours into the timecourse. A slight fall in the expression was observed between timepoints 8 and 12 after which, there was a gradual increase in expression until 20 hours. Subsequently, the rate of increase fell slightly although a moderate increase was observed between timepoints 20 and 24. Abundance decreased between 24 and 28 hours into the timecourse. Timepoint 0 expression was elevated in comparison to 24 and the same observation was made between timepoints 4 and 28.

Cluster (d) transcripts were at a mid-level of abundance at timepoint 0, increasing slightly between then and 4 hours before rapidly increasing to a peak 8 hours after theoretical dawn. Transcription levels remained almost constant in cluster (d) transcripts between 8 and 12 hours into the timecourse but fell rapidly by timepoint 16 hours. There was a gradual decrease in expression from 16 to 24 hours and a slight increase at 28 hours into the timecourse. Level of abundance was similar at theoretical dawn to the level of abundance 24 hours later, which was also the case when comparing timepoints 4 and 28 hours.

Amplitude of expression in transcripts under LL conditions was damped compared to transcripts under LD (Figure 19). Overall, two patterns (in clusters (a) and (b)) were observed showing transcripts that were at the peak of their expression a few hours subsequent to theoretical dawn, although cluster (a) peaked 4 hours later and

counts remained elevated for a period longer than in (b). The remaining two clusters, (c) and (d) grouped together transcripts that reached their highest expression levels during hours of theoretical nightime, although expression began accumulating again.

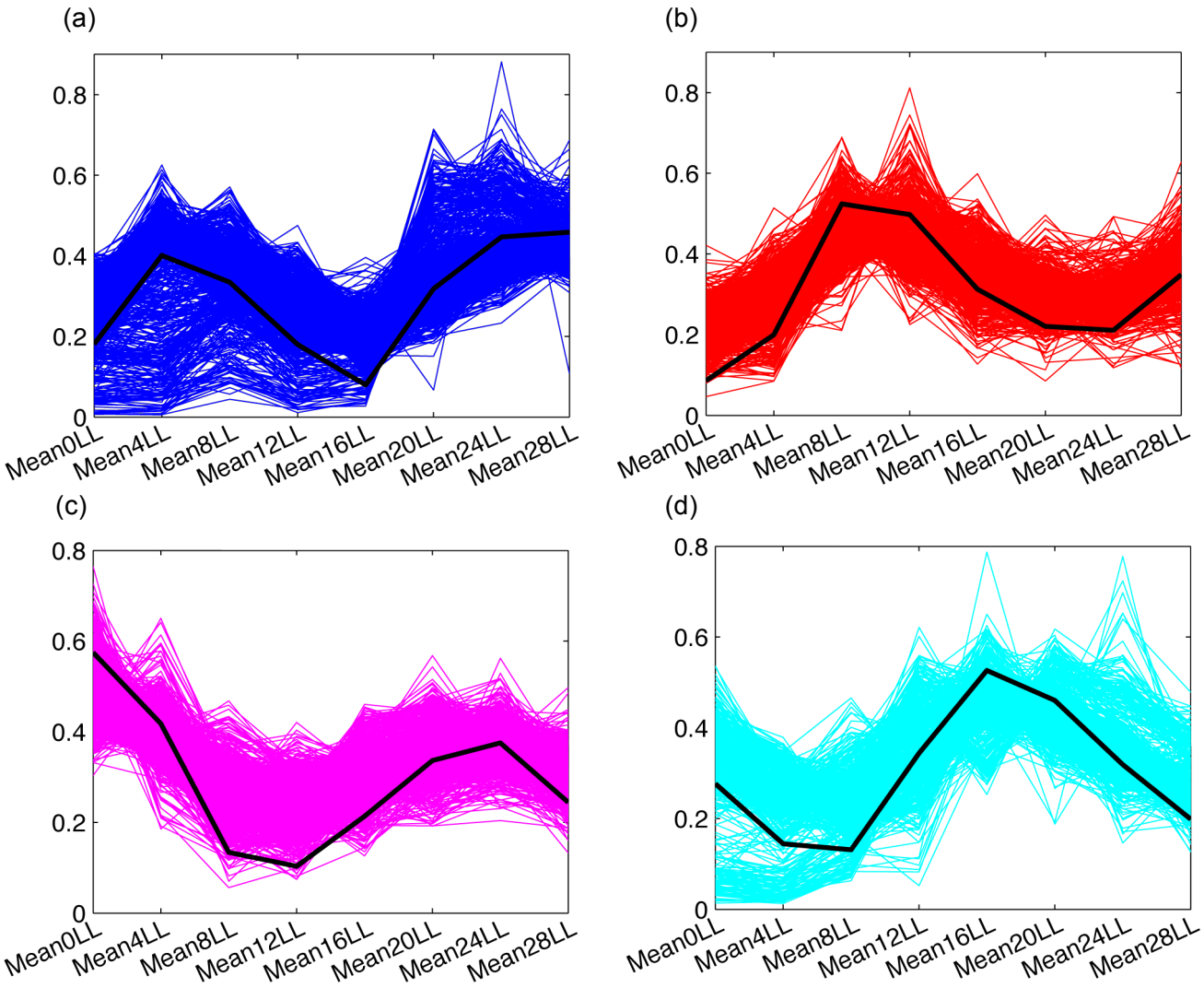


Figure 19 *K*-means clustering of transcripts under LL conditions
The normalised counts of transcripts under LD conditions clustered together according to expression profile similarity by association with the nearest mean. Data are standardised to fit an arbitrary scale on the y axis. The trend of each pattern is represented by a black line.

5.3.6.2 Pearson correlation of expression profiles

The expression profiles of 2,322 transcripts with temporal DE (padj≤0.001) under both LD and LL conditions were compared in all possible permutations. For example, BbPRR from LD was compared with itself, BbPRR from the LL dataset and all other transcripts under both conditions. 4,644 Pearson correlation values were calculated; one for each condition per transcript. Transcripts were clustered according to the distance metrics of the Pearson correlation values and double plotted in a heatmap of correlation values, ordered within their clusters.

The diagonal red line traversing the graph from top right to bottom left represents the perfect correlation of each transcript to itself under the same condition. Five large blocks of transcripts with positively correlated expression profiles were observed diagonally from top right to bottom left, with more strongly correlated transcripts in the bottom left. The large blocks of transcripts with similar expression profiles encompassed smaller areas of transcripts with more highly correlated expression patterns as well as subsets that were less correlated.

Examining pairwise correlations between expression profiles of the transcripts revealed 4 groups of transcripts, which were in anti-correlation with each other, visualised as blocks of blue traversing diagonally from top left to bottom right. In the bottom left corner of the heatmap are two clusters of transcripts, which are adjacently situated by hierarchical clustering, however they are in anti-correlation. Aside from the anti-regulated groups of transcripts that were clustered distantly from each other, there were 2 groups of anti-regulated transcripts that clustered near to each other, observed straddling the vertical and horizontal midlines of the heatmap to the right of the distinct blue zone created by distantly clustered transcripts.

A small group of transcripts that clustered together were observed to have expression patterns uncorrelated to most other clusters of transcripts, visualised traversing the top of the heatmap and from top to bottom on the right- hand axis. No other cluster of transcripts displayed this characteristic.

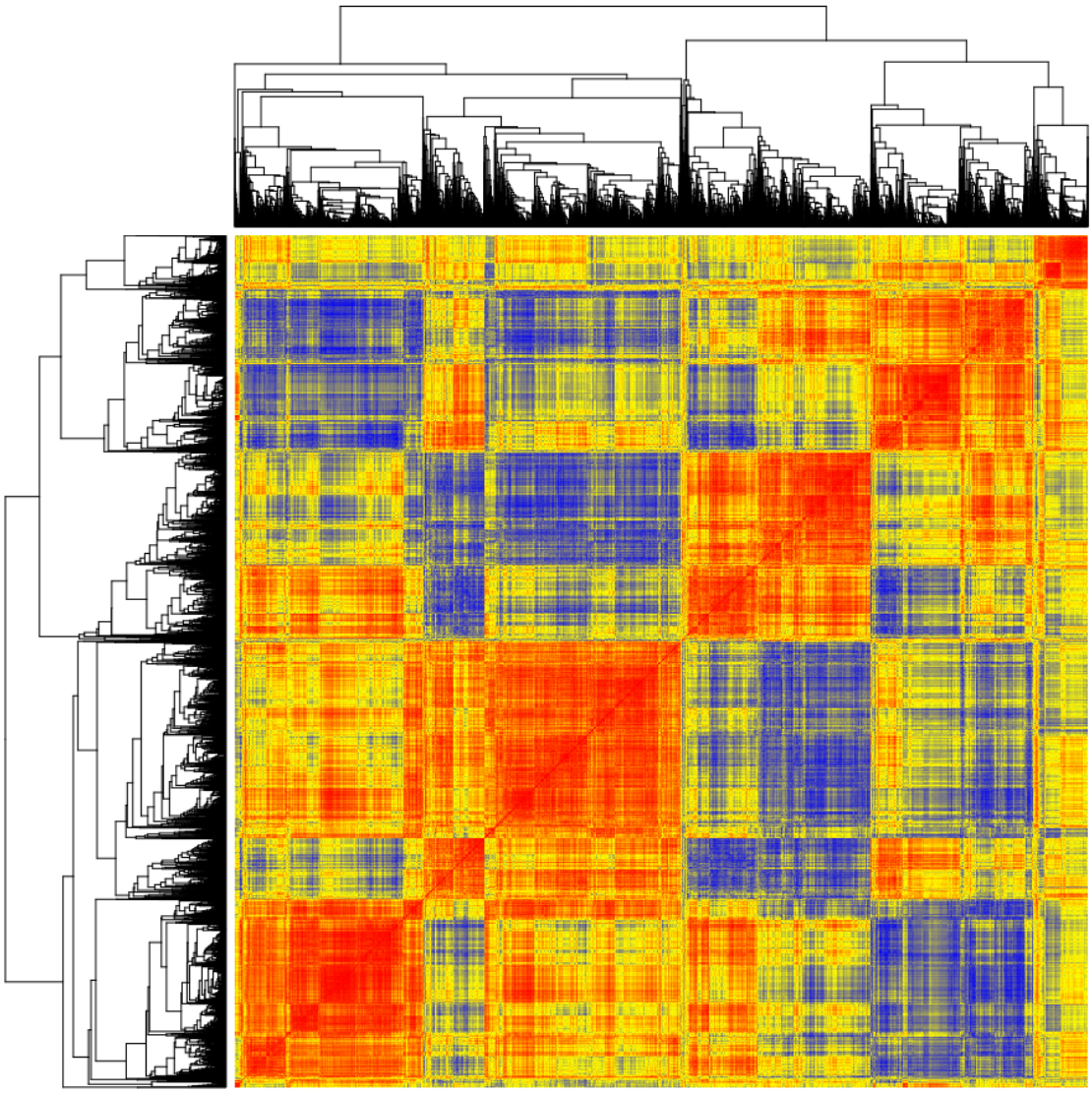


Figure 20 Clustergram of transcript expression

B. braunii transcripts were ordered by clustering (left and top axis) according to the distance metric of Pearson correlation values calculated of expression profiles in LD and LL. Pairwise comparisons were made between transcripts with temporal DE (padj≤0.001). Colour corresponds to correlation values ranging from -1 (blue, indicating negative correlation) to +1 (red, indicating positive correlation). Yellow represents values near to 0, indicating no correlation. Each transcripts is double-plotted so that the solid red diagonal line represents self-to-self comparison of transcripts.

5.4 Discussion

5.4.1 Contextualisation of bioinformatics methods

5.4.1.1 Read alignment using Bowtie

In this chapter, sequencing reads were aligned to the *de novo* reference transcriptome generated from the amalgamated data. As such, this reference represents all the *B. braunii* genes expressed over a 24 h diel cycle. In cases like this, where the exact reference sequence is available as opposed to that of a related species, a Burrows- Wheeler method of alignment is preferred, hence the use of Bowtie, a tool designed for the rapid alignment of many short reads between 4 and 1,024 bp, to a reference sequence where the average read quality is high, the number of reads per alignment is low and there are many successful alignments. Furthermore, when aligning to a transcriptome and quantification is more important than isoform detection, unspliced aligners, like Bowtie are better suited (Garber *et al*., 2011). In this study, the quality of the sequencing reads was exceptionally high, as demonstrated by an average Q_{PHRED} of over 30 (Table 1 section 3.3.2) for 100 bp reads. Lastly, computing time and memory costs of alternatives such as Maq and SOAP were prohibitive due to the size of dataset produced by this study but, thanks to the compression of the reference sequences by Burrows Wheeler Transform, Bowtie was fast and the memory footprint minimised.

5.4.1.2 Read quantification using HTSeq-count

The alignment file generated by Bowtie was then analysed using HTSeq-count to quantify the number of reads aligned to each transcript in the amalgamated reference from which an estimation of gene expression can be derived. HTSeq-count was preferred to generate this expression estimate data because it had the highest correlation with benchmark assays compared with the alternatives (RSEM, IsoEM and Cufflinks) (Chandramohan *et al*., 2013), reduced computational demand, and was relatively simple to use. Furthermore the HTSeq-count script was written specifically to provide count data for differential expression (DE) analysis and when the analysis was performed, no straightforward tool other than HTSeq-count dealt with ambiguous read mapping effectively. Since then alternatives such as *summarizeOverlap* in the *GenomicRanges* Bioconductor package and the C tool, *featureCount*, have been devised (Lawrence *et al*., 2013; Liao *et al*., 2014).

The paired-end design deployed in this work also made the use of HTSeq-count preferable, as the software can assimilate the duality of a paired-end read as a single

count; were it unable to do so, the read count would be inflated and potentially distort the expression estimates (Anders *et al*., 2014).

HTSeq-count provides expression estimates as raw counts per gene (or transcripts in this case), which was necessary as the chosen DE analysis package, DESeq2, uses count data instead of FPKM as an estimate of expression. Lastly HTSeq2 and DESeq2 were created by the same team, allowing for simple and rapid transition from expression estimation to DE analysis.

The default exon- union method of gene counting for HTSeq-count was used, where reads mapping to any of the exons of a gene are counted, instead of counting only reads that are mapped to constitutive exons of a gene- exon intersection method. Changes in expression may be incorrectly estimated by either method where multiple isoforms of a gene exist. In the case of the exon-union method; expression of alternative gene isoforms is underestimated. However the intersection method can result in reduced power of DE analysis, which was the accepted trade- off as DE analysis was the priority in this work (Garber *et al*., 2011).

At the time of writing, no straightforward tool other than HTSeq-count dealt with ambiguous read mapping effectively. Since then alternatives such as *summarizeOverlap* in the *GenomicRanges* Bioconductor package and the C tool, *featureCount*, have been devised (Lawrence *et al*., 2013; Liao *et al*., 2014).

5.4.1.3 Statistical analysis of differential expression

Most available software packages for the analysis of differential abundance of transcripts (*i.e.* differential gene expression, or DE) are unable to take into account the data from replicate samples. In this study, particular care was taken to generate meaningful, biological replicates. Hence, DESeq2 was selected as the analytical software because of its capacity to calculate the variance between replicate libraries and infer the correct statistical significance based on that variance.

A limitation of DESeq2 and other alternative DE analysis packages such as EdgeR and RSEM, is their reliance on pair-wise comparison of expression data to a single user-defined time-point. In DESeq2 the only method of testing whether treatment and time has an effect on DE of a transcript is to perform the likelihood ratio test (LRT). The LRT calculated for each transcript the difference in log likelihood between incorporating timepoint information versus disregarding timepoint information, testing whether each transcript was differentially expressed with time, at any point in the time series.

Uniquely, instead of pair-wise comparison with a single defined reference timepoint, a recently published approach to RNA-Seq time series analysis employs

hidden Markov models (HMMs) to characterise gene expression dynamics over a time course. A statistical expression trajectory index is generated, accounting for the dependency of expression patterns in later samples on those at earlier timepoints and is incorporated into an HMM for each gene (Oh *et al*., 2013). However, this method is complex and was unpublished at the time of analysis.

DESeq2 uses negative binomial distribution models to measure count variability between replicates. Some alternatives, such as DEGSeq and Myrna, use only Poisson distribution. Poisson distribution has been demonstrated to efficiently model read distribution between sequencing lanes when technical replicates only were compared. However the Poisson distribution has reduced ability to model over-dispersion of gene counts and is therefore less able to cope with biological variability (Langmead *et al*., 2010; Wang *et al*., 2008). This chapter uses data from biological replicates to infer DE and therefore a negative binomial model was used during analysis. The negative binomial method of modelling results in increased precision in dispersion estimates and therefore more power in detecting DE between experimental conditions (Robinson & Smyth, 2007).

Count normalisation is part of the DE analysis and required because some libraries may have been sequenced at a greater depth due to sequencing bias. In this context, RNA-Seq length bias is not important because the emphasis was on inter-sample gene comparison, not inter-gene comparison within samples.

Multiple testing correction is necessary in very large scale datasets such as the one generated here, because the number of false discoveries scales linearly with the number of tests of DE carried out. In this investigation, 287,421 transcripts were tested for DE by DESeq2. With an unadjusted p value of 0.01, one would expect 2,874 transcripts to be falsely deemed differentially expressed (*i.e.* false positives). The incidence of false positives was also minimised by the removal of outliers by Cooks Distance estimation prior to statistical testing, although we had to accept that this procedure may result in the elimination of some truly differentially expressed genes, most likely those that are present only once in one sample. Even though it might miss some differentially expressed genes, a low false positive rate was essential to the analysis because of the impossibility to confirm findings by follow- up experiments involving the direct genetic engineering of *B. braunii*. Hence a highly stringent adjusted p value cutoff of 0.01 was implemented. Furthermore, a confidence threshold of 0.01 yielded a computationally manageable dataset for downstream analyses.

5.4.1.4 RNA-Seq contamination and dataset integrity

During the analysis, the dataset was found to contain sequences originating from *C. glabrata*. This contamination was detected as a batch effect that was traced to a single day during which the affected cDNA libraries were prepared in the sequencing laboratory. This observation resulted in a review of standard operating procedures and the implementation of increased decontamination procedures within the sequencing laboratory. Following identification, the contaminating sequences were removed from the *B. braunii* dataset.

5.4.2 Differentially expressed transcripts

A larger proportion of *B. braunii* transcripts from culture under LD (50%) passed the controls of independent filtering (p value adjustment) than those from LL (20%) cultures, resulting in a discrepancy of 86,147 transcripts more in the LD dataset than the LL. The removal of more transcripts from the LL dataset could be explained by the damped expression profile of genes under constant light conditions which, due to the very stringent filtering criteria applied, may cause the software to reject a given transcript as being differentially expressed (Gaudana *et al*., 2013; Schaffer *et al*., 2001).

Approximately 90% of transcripts were not differentially expressed (*i.e.* were constitutively expressed) in both LD and LL light regimes. In *C. reinhardtii*, approximately 3% of the genome appears to be circadian regulated (Kucho *et al*., 2005), hence our estimate of 10% differentially expressed genes due to photoperiod and the circadian clock is not improbable (Kucho *et al*., 2005). In a microarray analysis of differential gene expression in *A. thaliana*, 6% of genes were identified to be clock regulated (Harmer, 2000). Contrastingly, it was more recently been proposed that approximately one third of the *A. thaliana* genome is regulated by the circadian clock (Covington *et al*., 2008; Hazen *et al*., 2009). This high level of clock-controlled genes compared to unicellular algae may be explained by numerous facets of the terrestrial angiosperm lifecycle. Higher plants have differentiated cells, tissues and organs performing contrasting functions, all of which need to be coordinated throughout growth and life-cycle transitions, such as flowering, requiring complex regulation of gene expression (Koltunow *et al*., 1990; Lepisto & Rintamaki, 2012; Love, 2004; Masucci *et al*., 1996). *B. braunii* is a colonial alga, but is essentially unicellular – therefore lacking any complexity that might arise from differentiated cell types, tissues or organs (Weiss *et al*., 2012; White, 1999) all of which require the “forward coordination” conferred by a complex and pervasive circadian system. Furthermore, *B. braunii* are less exposed to

rapid and extreme changes in temperature, concentration of salts and metals and of course, availability of water, than the sessile and terrestrial higher plants, which have adapted by development of robust stress response networks (Kobayashi *et al*., 2007; Nover, 2001). Indeed, circadian controlled genes are enriched within the higher plant phytohormone, cold and stress response pathways (Dodd *et al*., 2005; Hazen *et al*., 2009).

Mapping the transcriptome to the KEGG database enabled a rapid visualisation of the enzymatic pathways that were either predominantly constitutively expressed or differentially expressed. Multiple transcripts may map to a specific enzyme, due to redundancy in the KEGG labeling and the presence of multiple transcripts with similar enzymatic functions in the dataset; this particularity causes some degree of overlap between the diagrams, but it is striking that most pathways combine both differentially and constitutively expressed genes; hence metabolism appears to be controlled by critical enzymatic nodes, rather than by holistic regulation of entire pathways.

An example of this control may be illustrated by Carbon metabolism: Differentially expressed transcripts, notably those from algae cultured in LD, were enriched for sequences that mapped to KEGG pathways related to energy production and carbon metabolism. However, the photosynthetic and carbon fixation pathways also contained sequences for which expression appeared constitutive, supporting the view that some components of these pathways remain consistently transcribed, irrespective of temporal and photoperiodic effects. Carbon fixation involves the dark reactions of photosynthesis that do not require light to function. Consequently, light-dependent regulation (either via phototransduction or the circadian clock) may be superfluous to that pathway. Alternatively, the dark reactions are slow and despite Rubisco being one of the most abundant proteins, it is also one of the most inefficient enzymes in plants (Ellis, 2010; Feller *et al*., 2007; Whitney & Andrews, 2001), so in divorcing Carbon fixation from light- mediated control, it is possible that the alga maintains this critical system at full capacity to avoid any potential bottlenecks in C-assimilation (Goulard *et al*., 2004; Morris *et al*., 2000). This may be particularly important in *B. braunii*, which may have an increased requirement for carbon fixation for hydrocarbon synthesis (Schwender *et al*., 2004).

5.4.3 Expression of sequences involved in terpene biosynthesis

Botryococcenes are synthesised via the terpenoid and sesquiterpenoid and triterpenoid pathways. Di and tetra-methylated botryococcenes are the most abundantly synthesised type of botryococcene in Race B spp., along with other C_{31}-C_{34}

squalene derivatives (Niehaus *et al*., 2011). Terpenoids are structurally diverse and important to multivariate processes within plants. Examples include the role of sterols in regulation of cell membrane properties, carotenoids as light-harvesting and light-protecting pigments and their incorporation into the most abundant organic pigment; chlorophyll as a side-chain (Eisenreich *et al*., 2001).

B. braunii hydrocarbon synthesis is viewed as constant. However, *B. braunii* grows slowly compared to other, more commonly cultured *Chlorophyceae*, and hydrocarbons are typically analyzed at daily, if not weekly intervals ((Metzger *et al*., 1988; Yoshimura *et al*., 2013)). Temporal resolution is therefore lost.

The two precursor compounds required for triterpenoid biosynthesis are the isoprenoids; isopentyl diphosphate (IPP) and dimethylallyl diphophate (DMAP). Until relatively recently only one pathway was thought responsible for production of isoprenoid compounds; the cytosolic Mevalonate-dependent pathway. Stable isotope incorporation studies in eubacteria and plants revealed a second biosynthetic pathway for isoprenoid production via the plastidic MEP/DOXP pathway (Rohmer *et al*., 1993; Schwartz, 1994). Both the MEP/DOXP and Mevalonate-dependent pathways are present in the algal clade, although green algae thus far seem to exclusively utilise the MEP/DOXP pathway and *B. braunii* transcriptomics studies performed thus far agree (Eisenreich *et al*., 2001) (Molnár *et al*., 2012).

(2Z,6Z)-farnesyl diphosphate synthase and short-chain Z-isoprenyl diphosphate synthase are important components of the botryococcene synthesis pathway because they are responsible for conversion of IPP or DMAP into (2Z,6Z)-Farnesyl diphosphate and 2-cis,6-trans-Farnesyl diphosphate; (2Z,6E)-Farnesyl diphosphate respectively. Farnesyl-diphosphate farnesyltransferase then converts either (2Z,6Z)-Farnesyl diphosphate or (2Z,6E)-Farnesyl diphosphate into Presqualene diphosphate. Presqualene diphosphate is subsequently converted by farnesyl-diphosphate farnesyltransferase, otherwise known as squalene synthase, into Squalene. It has been demonstrated that C_{30} botryococcenes are synthesised via this same pathway and C_{30}- C_{34} botryococcenes are synthesised in a series of modifications downstream of this point, although the genetic and enzymatic basis of the catalysis of C_{32} to C_{34} botryococcenes is yet to be characterised (Molnár *et al*., 2012; Niehaus *et al*., 2012; 2011).

Enzymes mapped by transcripts that are temporally differentially expressed in both LD and LL conditions can be said to be under circadian control because oscillation in abundance is robust to constant conditions- a characteristic indicative of circadian clock components and their influence on transcription. Enzymes that are mapped to by transcripts that are temporally differentially expressed under LD

conditions but not LL can be thought of as under photoperiodic control due to the apparent requirement of environmental cues to mediate their transcription.

The *B. braunii* MEP/DOXP pathway for triterpenoid synthesis was nearly entirely completed by the mapping of transcripts that are under circadian control and several of the same enzymes could be alternatively mapped by transcripts under photoperiod control. Constitutively expressed transcripts have little involvement with the MEP/DOXP pathway. However the enzymes, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase and farnesyl diphosphate synthase were all identified in the constitutively expressed dataset. These enzymes were also mapped to by transcripts under circadian and photoperiodic controls. Further within the constitutively expressed dataset, the place of farnesyl diphosphate synthase in formation of Farnesyl PP, is taken by short chain z-isoprenyl diphosphate synthase.

Evident from these data gene redundancy may have a role in the regulation of botryococcene synthesis although the precise mechanism of this role cannot be determined on the basis of differential expression analysis alone. These data agree with similar findings of circadian control over components of the MEP pathway in higher plants (Covington *et al*., 2008).

Interestingly the differentially expressed transcripts for *B. braunii* cultured in LD map exclusively to the MEP/DOXP feed in route to the squalene synthesis pathway, However the non-differentially expressed dataset contained transcripts that mapped to three components of the mevalonate pathway and several critical nodes were left unmapped in the MEP/DOXP pathway.

Although from this analysis it appears that the terpenoid pathway in *B. braunii* is differentially regulated throughout the day, further analysis of the co-expression data to identify the specific transcripts involved is required. Unfortunately, lack of time precluded this analysis.

5.4.4 Expression profiles of the predicted circadian clock sequences

In chapter 4, a model of the *B. braunii* circadian clock was proposed, and consisted of a single loop of two interacting components; *BbPRR* is a homologue of the pseudo response regulators in *A. thaliana* and *O. tauri* (of which *TOC1* is the archetype), and therefore termed *BbPRR*; BbCCA1, a homologue of *CCA1/LHY*, termed *BbCCA1* (Matsushika *et al*., 2000; Wang *et al*., 2008). In this chapter, the expression profiles of each of these sequences were analyzed and compared to their counterparts in *O. tauri*, *A. thaliana* and *Oryza sativa*.

For algae cultured in a 12L:12D photoperiod, *BbPRR* expression is comparable to that of *O. tauri*, *A. thaliana* and *Oryza sativa*, in that there is a clear peak of expression towards the latter third of the photoperiod and a reduction in expression in the scotophase. Moreover, *BbPRR* expression increases slightly before dawn on the standardised diagram, which is indicative of a circadian component (Alabadi, 2001; Duffield *et al*., 2002). Compared to the data for the other organisms, *BbPRR* expression appears phase- shifted by approximately 2 h. However, the data for the model organisms used for comparison was collected at three- hourly intervals in contrast to the *B. braunii* data, which was collected at four- hourly intervals (Corellou *et al*., 2009; Hsu *et al*., 2013; Murakami *et al*., 2006). The observed phase- shift is therefore probably due to the lower level of temporal resolution achieved in this investigation.

BbCCA1 expression in algae cultured in a 12L:12D photoperiod is seemingly in anti- phase compared to that of its homologues in *A. thaliana* and *Oryza sativa*, in that peak expression occurs at dusk instead of just prior to, or at dawn. However, *O. tauri CCA1* peaks in expression in the middle of the scotophase, just four hours later than the observed peak of *BbCCA1*. Consequently, it is quite possible that, with higher temporal resolution, we could assess whether the algal *CCA1*s possess more similar expression profiles than *BbCCA1* and higher plant *CCA1*s (Corellou *et al*., 2009; Hsu *et al*., 2013; Murakami *et al*., 2006). Again, the observed phase shift may be compounded by differences in temporal resolution between studies of model organisms and that achieved in this investigation.

When cultured in LL, the expression levels of *BbPRR* shows no cycling, suggesting that *BbPRR* is light-responsive or at least sensitive to light at all points in the photophase. The expression of *BbCCA1* is damped in LL, and, as for LD, is in antiphase to *AtCCA1* and *OsCCA1* expression. However, there is no available data for *O. tauri*, and it is therefore not possible to compare *BbCCA1* expression to that of a homologue in algae. Taking into account the more limited temporal resolutions in this investigation, these data agree with the model proposed for the *O. tauri* clock (Troein *et al*., 2011). In this scenario, light induces *TOC1* and *CCA1* transcription, TOC1 positively regulates *CCA1* transcription and CCA1 negatively regulates *TOC1* transcription. The constitutive expression of *BbPRR* in algae cultured under LL is at a level close to peak expression observed in *B. braunii* under LD. On the other hand, the maintenance of an oscillatory expression profile of *BbCCA1* in algae cultured in LL suggests that light stimulus is temporally gated for this component.

5.4.5 Co- expression analysis of the differentially expressed transcripts

Gene families typically have similar temporal gene expression profiles. However investigation of genes with no known functional relationship that are co- expressed can lead to characterisation of associations between genes and pathways of interest. Cluster analysis was used to elucidate networks of co- and anti- expression in the *B. braunii* transcriptome. Both hierarchical and non-hierarchical cluster analysis was performed using only transcripts with padj ≤ 0.001, as the computation and visualisation of all possible pairwise comparisons between expression profiles of transcripts using a cutoff of padj ≤ 0.01 was unfeasible due to the large scale of the dataset.

Hierarchical clustering was performed to group together networks of co- and anti-expressed transcripts in the whole dataset, *i.e.* comparing gene expression from algae cultured in LL and LD Figure 18 and Figure 19.

Hierarchical clustering groups genes according to indices of pairwise similarity, which are calculated from the expression data. For these data, Pearson correlation metrics were used. Subsequently relationships between the clusters are analyzed, building a dendrogram that represents a hierarchy of genes in clusters that are grouped into higher- level systems (Eisen *et al.*, 1998; Kuklin *et al*., 2002). Performing clustering in this way is quantitative and provided a means of extracting transcripts that were clustered together using the calculated Pearson metrics as a cutoff. However, the interpretation of hierarchical clustering results can lead to false grouping of transcripts as the solution is based on a single pass of the data; there are no re-evaluations of placement once a data point is assigned to a node and early incorrect assignments are not re-adjusted (Tamayo *et al*., 1999). This method revealed several large networks of co-expression and anti-expression within the differentially expressed dataset: A group of transcripts were observed that displayed neither co- expression or anti- expression with any other cluster may be an interesting avenue to explore, as they are seemingly disconnected to what seems to be a highly complex network of expression comprising the *B. braunii* transcriptome. A suggested explanation for their independent behaviour may be that they are the result of residual contamination originating from *B. braunii* non- axenic cultures or during sequencing library preparation.

Non- hierarchical clustering was therefore performed on the separate datasets from algae cultured in LL and LD to resolve expression patterns of the co- expressed transcripts (Figure 18) and (Figure 19).

The *K*-means method of non- hierarchical clustering involves an iterative process of re- adjusting the placement of data points in clusters until the optimum solution is found, and transcripts with similar temporal expression profiles are

partitioned into clusters with the nearest mean (Ihmels *et al*., 2003; Kharchenko *et al*., 2005; Kotlyar *et al*., 2002). This method enabled the transcripts of the *B. braunii* transcriptome to be grouped into four clusters that represented visualisation of the expression profiles. The limitation of cluster analysis using *K*-means is that it is not quantitative as is the case with hierarchical clustering. Additionally, the number of clusters is user specified and therefore prior knowledge of dataset characteristics is important for generation of optimal cluster solutions (Quackenbush, 2001). Scree plots were used to determine how many clusters to specify for *K*-means cluster analysis by demonstrating how many patterns account for the majority of variance in the temporal expression profiles. A four- cluster solution was chosen because the position of the elbow in the Scree plots for both 12:12 photoperiod and continuous light datasets suggested that approximately between 70% and 80% of the variance in temporal expression pattern could be represented by the clustering of transcripts into four groups, each with its own common pattern (see Appendix Figures 23 and 24). Considering the *K*-means data from both LL and LD, four distinct expression patterns were observed in the expression data of the *B. braunii* transcriptome. However in comparison to the patterns generated by *B. braunii* expression under LD, the LL profiles were damped. Phase- advancement was not apparent in the LL clustered expression profiles as has been observed in other Eukaryotes exposed to constant light (Knoch *et al*., 2004; Matsushika *et al*., 2000). However, these observations in previous studies were made of circadian genes in targeted and prolonged timecourses and therefore this investigation may not have captured ensuing phase- shifts in the expression patterns.

By themselves, the *K*-means and the hierarchical cluster analyses clearly show major networks of co-regulated gene expression. However, further mining of the data with respect to specific gene sequences and function is required to better understand the data. Individual characterisation of functional domains, as well as KEGG pathways mapping of the transcripts within the clusters of expression patterns would give insights into the phasing of genes of interest, metabolic pathways and regulation of cellular processes. *K*-means analysis can provide further insight into networks of co-expression by revealing whether these relationships are maintained under both conditions of LD and LL. Finally, the calculation of Pearson correlation metrics provided a database of quantitatively correlated transcripts that were co- expressed and anti-expressed, also representing a useful tool for investigation of pathways of interest in *B. braunii*.

5.5 Summary

In this chapter, temporal differential expression analysis of the annotated *B. braunii* diel transcriptome was performed using raw count data generated by time-series RNA-Seq. *B. braunii* transcripts with differential, non-differential and light-dependent differential expression were mapped to the higher plant circadian clock circuit, the biosynthetic pathway of Botryococcene synthesis and the KEGG overview of metabolic pathways. Mapping of transcripts to the MEP/DOXP pathway for Botryococcene synthesis was in agreement with existing biochemical data and demonstrated photoperiodic control of critical node components within this pathway, suggesting photoperiodic control of hydrocarbon biosynthesis in *B. braunii,* which may be ascribed to temporal resource or metabolite partitioning. Using both normalised count data and standardisation of count data the expression profiles of *B. braunii* predicted clock components were compared to those of model organisms, revealing similarities in expression profiles particularly under LD conditions although phase shifts were observed. Finally, cluster analysis of expression profiles of the most significantly differentially expressed *B. braunii* transcripts was performed, comparing transcripts under both conditions of LD and LL and subsequently within their respective conditions. LL and LD co-regulated transcripts were predominantly clustered within five groups and anti-regulated mainly within four groups. Four cluster solutions accounted for the majority of all expression profile pattern variability in LD and LL conditions when analyzed independently. Networks of co-expressed transcripts from both cluster analyses were exported to be accessible for future work.

DISCUSSION

6.1 The potential of Next Generation Sequencing for investigating microalgal physiology

Comprehensive knowledge of entire genomes has enabled understanding of fundamental biological processes at the whole-organism or "systems" level (Chow & Kay, 2013; Cushman & Bohnert, 2014; Urano *et al*., 2010). A complete and annotated genome serves as the reference for other holistic technologies including transcriptomics, proteomics metabolomics. To date, there are 12 publically available (and more or less well annotated) genomes from green microalgae, including those of *Chlamydomonas reinhardtii*, *Chlorella variablis*, *C. vulgaris*, *Coccomyxa subellipsoidia*, *Dunaliella tertiolecta*, *Volvox carterii, Micromonas pusilla, Micromonas SP. R299 and Helicosporidium* sp. ATCC 50920, *Ostreococcus tauri*, *O. lucimarinus*, *Auxenochlorella protothecoides* (Blanc *et al*., 2012; 2010; Ferraz *et al*., 2006; Gao *et al*., 2014; Guarnieri *et al*., 2011; Merchant *et al*., 2007; Pombert *et al*., 2014; Prochnik *et al*., 2010; Rismani-Yazdi *et al*., 2011; Worden *et al*., 2009). Mining this genomic data has revealed a variety of biosynthetic gene clusters, that have the potential to be industrially relevant, such as those involved in carbon sequestration and fixation, anaerobic fermentation and hydrogen metabolism and synthesis of triacyl- glycerols, selanoproteins, carotenoids and vitamins (Cordero *et al*., 2012; Grossman *et al*., 2007; Merchant *et al*., 2007; Radakovits *et al*., 2012). These micro-algal species were sequenced mainly because of the possibility of generating axenic cultures from which the algal DNA is purified, thereby facilitating both the sequencing and annotation. However, some, possibly most, microalgae cannot be easily disassociated from co-cultured microbes (Lee *et al*., 2013; Rivas *et al*., 2010), thereby limiting the speed with which algal genomes can be generated, due to high levels of contaminating bacterial DNA. For this reason, there is currently no assembled and annotated genome publically available for *Botryococcus braunii*. However, the DOE Joint Genome Institute commenced the sequencing of the *B. braunii* race B strain "Showa" genome in 2010 (NCBI BioProject ID 60039).

Fortunately, due to the development of massively parallel, short-read sequencing ("Next Generation Sequencing; NGS) and *de novo* sequence assemblers, prior knowledge of genomes is no longer a prerequisite for elaborate systems-based analyses. While transcriptomic studies of microalgae have typically been performed

using microarrays designed for "model" microalgae, most notably *Chlamydomonas* (Kim *et al*., 2011; Nguyen *et al*., 2008; Toepel *et al*., 2011), next generation RNA sequencing (RNA-seq) is not restricted to heterologous comparisons and reports a larger dynamic range of transcript levels. Hence RNA-seq is increasingly employed in the elucidation of molecular pathways and gene regulatory networks (Bochenek *et al*., 2013; Gonzalez-Ballester *et al*., 2010; Rismani-Yazdi *et al*., 2011). The "post-genomics" technologies have therefore enabled detailed investigation of the dynamics of gene expression in microalgae, allowing transcriptional controls of genes and biosynthetic gene clusters to be elucidated (Miller *et al*., 2010; Yang *et al*., 2013).

RNA-Seq provides an accurate and comprehensive archive of sequence data that together describes exactly what pathways are active and the biological processes that are being undertaken within an organism at a given time point. The sequence information may be used for both qualitative and quantitative analysis following comprehensive annotation and differential expression analysis. The quantitative analysis of transcripts in time series RNA-Seq data can be used in the identification of genes of interest, for which expression profiles are already known or when comparisons can be drawn from homologues in model organisms. The primary aims of this project were therefore to generate and analyse mRNAseq-derived data to first, elucidate a molecular basis for the *B. braunii* circadian clock, second, to identify the molecular basis of the metabolic pathway for Botryococcene synthesis and, third, to determine whether this hydrocarbon production pathway was differentially regulated over a daily cycle.

6.2 Sequencing, annotation and analysis of the *Botryococcus braunii* transcriptome

In this project, the dynamic, diel transcriptome of *Botryococcus braunii* was sequenced, assembled, annotated and mined for salient functions including the circadian clock network and the molecular pathway for hydrocarbon biosynthesis. Assembly of a transcriptome without a reference genome sequence remains a new and challenging field. Obstacles include the overall number of sequenced reads, the range of concentrations of the RNA molecules that generated these reads, the complexity imparted by alternative splicing, allelic variation and gene duplication, and the detection of sequencing errors (non-random and random) (Li *et al*., 2014; Tulin et al., 2013).

De novo transcriptome assemblers, such as that used in this study (the Trinity pipeline) have user-tuneable parameters associated to deal with these complexities.

However, if little is known of the studied organism, often the case when no genome is available, tuning these parameters and interpretation of data is complex. In this project, the assembly of the *B. braunii* transcriptome was particularly challenging because of the lack of a reference genome and the presence of bacterial and fungal sequences in the raw data. Axenic culture of *B. braunii* cannot be successfully maintained and bacteria become strongly associated with the extracellular matrix of colonies such that culture contamination can be minimised but not eradicated. This is the primary reason that *B. braunii* genome assembly has not thus far been achieved. In this study, fungal contamination issues also arose from within the sequencing facility.

Studies of the *B. braunii* race B transcriptome have previously been published (Ioki *et al*., 2012; Molnár *et al*., 2012). However, both published studies included only one replicate. The Molnar study sampled 3 timepoints at intervals spread over a period of 5 days, and, regardless, the RNA was pooled into just one sample prior to sequencing. The Ioki study sequenced the RNA of only one culture at one timepoint. Consequently, whilst gene abundance was ascribed in both earlier studies, no differential expression analysis or diel expression profiling was performed. The *B. braunii* transcriptome presented here is, in sharp contrast to these previous investigations, the most complete and in depth to date. Two thousand times more sequencing data were generated than the largest previous study from Molnar (2,011,057,142 reads compared to 1,334,609). Moreover, the transcripts generated in this study are, on average, longer than those in previous investigations; 1,014 bp compared to 746 bp. Prior to analysis the majority of transcripts that did not originate from *B. braunii* were removed from the assembly generated in this study, whereas potential contamination was retained in previous, published transcriptomes of *B. braunii*. Most importantly, the transcriptome was derived from algae that were harvested over an entire diel cycle, not just at a single timepoint per day, providing both coverage of differentially expressed genes and crucial temporal information that amounts to the dynamic transcriptome of *B. braunii*. Finally, all possible alternative transcript isoforms were reported.

The distribution of transcript annotations within the GO category “Molecular Function” demonstrates a comprehensive inventory of *B. braunii* transcriptional activity. Knowledge of molecular function is crucial for the possible application of genetic engineering technologies to *B. braunii*, or the translation of *B. braunii* molecular pathways to alternative hosts as performed in other studies (León *et al*., 2007). Further, when considering KO assignment alone, 59% of transcripts were assigned a functional annotation, an improvement on prior transcriptome studies of *B. braunii* where 45% of transcripts were annotated with functional annotation (Molnár *et al*., 2012). The

distribution of annotations was similar to that in the Showa transcriptome, with high proportions of transcripts assigned to pathways involved in the metabolism of carbohydrate and energy overall, and the biosynthesis of lipids, nucleotides, amino acids, cofactors and vitamins in both.

The annotated (protein or RNA) sequence data generated in this study will be submitted to the NCBI Genbank database and made publically available via a web-based portal (http://www.ncbi.nlm.nih.gov/genbank/), facilitating further exploration of *B. braunii* omics and informing future genetic studies of the alga. Furthermore, we have described herein a bioinformatic workflow for the *de novo* assembly, annotation and analysis of RNA-seq data from non- model eukaryotic organisms, which can be applied to the transcriptomics study of other lesser- characterised organisms.

6.3 Identification and characterisation of circadian clock components of the *Botryococcus braunii* transcriptome

Chapter 4 addresses how the annotated transcriptome of *B. braunii* was mined for circadian clock components and the development of a clock model comprising two components. The two *B. braunii* transcripts that were incorporated into the proposed circadian clock model showed good sequence homology, shared conserved domains and sequence motifs with homologous clock proteins from other plants, and were consequently termed BbPRR and BbCCA1. The predicted *B. braunii* clock components are orthologous to those found in higher plants and green algae, although the proposed model is representative of just the very core of the higher plant clock model. The *B. braunii* clock model was in agreement with the only other fully described clock model for green algae; that of *Ostreococcus tauri*. The *O. tauri* clock model is massively reduced compared to that of higher plants, and comprises only a TOC1 component and a CCA1/LHY-like component, both orthologs of their counterparts in *A. thaliana* (Thommen *et al*., 2010, Corellou *et al*., 2009).

The identification of *BbPRR* and *BbCCA1* as homologues to the angiosperm and *O. tauri* core clock components, *TOC1*, the *PRR*s and *CCA1/LHY* was supported by the presence of *B. braunii* transcripts homologous to peripheral components of the higher plant clock, *ELF3*, *ELF4* and *LUX*. This is a novel discovery, as *ELF3* and *ELF4* have only ever been found in land plants (Locke *et al*., 2006). The *B. braunii* sequence homologues of *ELF3* and *ELF4* were not named *BbELF3* and *BbELF4*- this would have been somewhat presumptive as after identification by sequence homology they were not further characterised. However after further investigation, such as that performed for *BbPRR* and *BbCCA1,* the *B. braunii* sequence homologues of *ELF3* and *ELF4* may

be confirmed as circadian components and confidently renamed. This data would provide further elucidation of the *B. braunii* circadian clock and further characterise evolutionary relationships within the Chlorophyta by discovery of a clock mechanism bridging the gap between that of *O. tauri* and that of higher plants.

In Chapter 5 temporal gene expression analysis of the annotated *B. braunii* diel transcriptome was performed. Temporal expression patterns of *B. braunii* predicted clock components were compared with model counterparts, showing similar waveforms of oscillatory expression. Advanced phasing of peak expression was observed when comparing the *B. braunii* predicted clock components, *BbPRR* and *BbCCA1* with their counterparts in model organisms. The phasing of *BbPRR* and *BbCCA1* was closer to that of its counterpart in *O. tauri* than in higher plants and like *O. tauri* created a wider waveform than the profiles of the components from higher plants. It was not possible to compare *BbPRR and BbCCA1* mRNA expression with their *O. tauri* counterparts under constant light conditions as the data for *O. tauri* were not available. However generation of transcriptional and translational luciferase reporter lines in demonstrated that CCA1 and TOC1 circadian oscillation was maintained, although somewhat damped, in *O. tauri* cultured in constant light (Corellou *et al*., 2009). Given these data, it seems probable that sampling at shorter intervals may reveal further similarites between the clock component profiles of *B. braunii* and *O. tauri*.

The longer duration of peak expression in algal clock components may be due to the reduced need for fine-tuned output from the simple clock model. Higher plants have adapted to the rapid flux of their terrestrial environment via an intricate network of regulation that allows anticipation of or rapid response to sudden and extreme changes in temperature, concentration of salts and metals and of course, and availability of water. It seems probable that the increased complexity of the higher plant clock reflect the requirement to modulate these robust stress response networks require (Dodd *et al*., 2005; Hazen *et al*., 2009; Kobayashi *et al*., 2007; Nover, 2001). It is perhaps the case that more defined expression (*i.e.* narrower waveforms are generated by peaks) of numerous clock components impart the necessary flexibility and fine- tuned response. Although, broadly speaking, the higher degree of similarity between *B. braunii* and *O. tauri* clock components is unsurprising giving the close evolutionary relationship (Hindle *et al*., 2014; Keeling *et al*., 2005).

In the future, should assembled genomic data become available *B. braunii,* predicted *PRR* clock components could be further confirmed using computational methods such as BLAST, to search for the conserved Evening Element (AAAATATCT), known to regulate *TOC1* transcription in *A. thaliana* and *O. tauri* (Alabadi, 2001; Corellou *et al*., 2009). Alternatively, inverse polymerase chain reaction

followed by product sequencing could be used to identify promoter region sequences of the predicted clock components (Triglia *et al*., 1988). Unfortunately no reliable molecular toolkit for the transformation of a suitable algal host is available to empirically confirm and further define the roles of *BbPRR* and *BbCCA1*. However, the use of *A. thaliana* in a complementation study provides a suitable alternative and is an avenue to explore in future work. Other *B. braunii* predicted proteins were identified in this investigation, by a search of the *B. braunii* open reading frame database with custom HMMs generated from model clock proteins. These *B. braunii* proteins, similar to characterised clock proteins provide a reservoir of potential clock candidates to be mined for other components involved in *B. braunii* circadian rhythmicity.

Further considerations are that a longer period of sampling for RNA-seq would have allowed more confidence to be imparted in the maintenance of oscillation in transcription of predicted *B. braunii* clock transcripts, as other published work has described expression profiling for up to 72 hours. Furthermore, a shorter sampling interval would allow expression patterns of predicted clock genes to be elucidated to a finer resolution, allowing a more direct comparison between *B. braunii* transcripts and those of model organisms, which are more often sampled at 3 hourly intervals than 4 (Corellou *et al*., 2009; Hsu *et al*., 2013; Matsushika *et al*., 2000; Murakami *et al*., 2006). However, it was not feasible to perform a study of greater duration and resolution single- handed and whilst maintaining the degree of replication and considering the computational requirements of analysis.

6.4 Mapping and characterisation of the *Botryococcus braunii* triterpenoid synthesis pathway

Differentially expressed, constitutively expressed or transcripts with light dependent differential expression were mapped to the KEGG metabolic pathways overview, the plant circadian clock circuit and the terpenoid pathway. The identification and mapping of temporally differentially expressed *B. braunii* transcripts to the KEGG terpenoid pathway not only confirms that precursors for botryococcene synthesis are produced via the MEP/DOXP pathway, but also demonstrates that this process is largely under photoperiodic and/ or circadian control (Eisenreich *et al*., 2001; Molnár *et al*., 2012). However, 3 enzymes were also identified in the constitutively expressed dataset, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase and farnesyl diphosphate synthase. It is evident from the distribution of enzyme identifications amongst transcripts that are temporally regulated and those that are constitutively expressed that there is a degree of genetic

redundancy in the *B. braunii* transcriptome although the precise purpose or mechanism of this regulation remains elusive.

The genetic and enzymatic basis of the downstream modification of C_{30} botryococcenes to C_{34} botryococcenes was not elucidated by annotation performed thus far on the *B. braunii* transcriptome. C_{34} botryococcenes are synthesised in a series of modifications downstream of IPP or DMAP conversion to farnesyl diphosphate via the MEP/DOXP pathway and subsequent production of C_{30} botryococcene via squalene synthase. Previous studies have identified the role of Adomet- dependent methyl transferase in modification of C_{30} to C_{32} botryococcenes by targeted searches of *B. braunii* transcriptomics data (Molnár *et al*., 2012; Niehaus *et al*., 2012; 2011). Time constraints precluded the opportunity to perform similar searches. However targeted searches for methyl-transferase enzymes involved in botryococcene should be included in future work, particularly considering the increased likelihood of identification imparted by the more comprehensive nature of the transcriptome presented in this investigation. Computational methods, such as BLAST searches using known methyl-transferase sequences as queries, and screening of the predicted ORF database using Hidden Markov Models generated from known methyl transferases could be employed to provide a shortlist of candidates. Subsequent insertional mutagenesis studies in a host organism such as *C. reinhardtii* or a model yeast strain could be employed to confirm biochemical function of the shortlisted candidate genes. Fortuitously, protocols have recently been developed for use of *C. reinhardtii* as a platform for the identification of insertional mutants without the need for a selectable phenotype assay, which was a limitation of metabolic pathway characterisation by reverse genetic approaches (Dent, 2005; Gonzalez-Ballester *et al*., 2011). In contrast with previous studies, future work to identify methyltransferases involved in botryococcene modification should not be limited to the Adomet- dependent group, as methylation of C_{32} to C_{34} botryococcenes could not be accounted for by this target group of enzymes alone (Niehaus *et al*., 2012).

6.5 Regulatory networks of gene expression in the *Botryococcus braunii* transcriptome

The current view of *B. braunii* hydrocarbon synthesis is that it is constant throughout the diel cycle. The *B. braunii* growth rate is slow compared to other, more commonly cultured *Chlorophyceae* (Largeau *et al*., 1980; Metzger *et al*., 1985), and temporal fluctuations in hydrocarbon synthesis have not previously been resolved as

analysis has typicallybeen performed on a daily, or even weekly basis (Metzger *et al*., 1988; Yoshimura *et al*., 2013).

From the analyses of data produced in this investigation, it is apparent that botryococcene production is differentially regulated throughout the day but it was not possible to clarify the nature of that temporal regulation due to time constraints. Further mining of co- expression data is required to elucidate the precise mechanisms by which this regulation occurs, *e.g*. via direct temporal gene regulation or indirectly via metabolite partitioning.

Knowledge of metabolite partitioning is critical to commercial production of biofuels from microalgae. A major limitation of the "photosynthesis- to- fuels" pipeline The preference of photosynthetic organisms to direct carbon resources towards sugar biosynthesis leading to biomass accumulation, instead of the terpenoid and fatty acid biosynthetic pathways, which are typically allocated 5 and 10% of the carbon pool respectively (Melis, 2013). The B race of *B. braunii* is unusual in that it partitions carbon in favour of biofuels production with terpenoid and fatty acid synthesis monopolising carbon flux (45% and 10% respectively) (Eroglu & Melis, 2010; Eroglu *et al*., 2010). However, this leads to slow biomass accumulation. A more in depth knowledge of the regulations over resource partitioning and other rate limiting factors in pathways related to photosynthesis, carbon fixation and terpenoid and fatty acid synthesis pathways may enable the optimum balance of hydrocarbon production and biomass accumulation to be attained in *B. braunii* and other candidates for biofuel production.

Future analysis of the temporal gene expression profile of the whole botryococcene synthesis pathway using data generated in this investigation would elucidate interactions between the genes involved in the pathway, the times of day at which the rate of production is greatest and when the highest quality *i.e.* the most methylated botryococcenes are produced. Hypotheses formulated based on this data can be empirically tested by targeted specific timepoints in the day to extract and analyse hydrocarbon content of *B. braunii*, using GC-MS or Raman Spectroscopy. Circadian and diurnal controls over metabolic pathways potentially have important implications for the industrial applications of *B. braunii* because information in this area would allow the timing of oil extraction and downstream processing methods and to be optimised. Furthermore, from a physiological perspective, a deeper understanding of resource allocation and partitioning in *B. braunii*, which may be extended to other single-celled or colonial members of the *Chlorophyta* would be enabled.

Analysis of transcriptomics, metabolomic and genomic data have in recent years revealed elaborate regulatory and signaling networks of global gene expression,

protein modification and metabolite composition in several model plants and algae (Cushman & Bohnert, 2014; Grossman *et al*., 2003; Jain *et al*., 2007; Ramos *et al*., 2011; Shrager, 2003; Urano *et al*., 2010). Cluster analyses of the *B. braunii* transcriptome revealed multiple clusters of transcripts with different patterns of co- or anti- expression over a 28 hour cycle that were robust to conditions of constant light. These findings suggest circadian regulation of a complex network of genes and molecular pathways in *B. braunii*, as in other Chlorophyta, despite the architectural simplicity of the proposed clock model. In recent years, integrated analyses of transcriptomic and metabolomic data has demonstrated connections in networks of gene expression and metabolites, elucidating complex regulatory networks, largely those involved in abiotic stress in crop plants in the hope of increasing yields (Armengaud *et al*., 2009; Maruyama *et al*., 2009). Furthermore, connections between circadian clock regulation and metabolite production have been revealed (Nakamichi, 2011). The integration of the transcriptomic dataset presented here with data resulting from the metabolic profiling of *B. braunii* could shed light on the regulatory networks represented by the clusters of gene expression observed, with the similar goal of using this knowledge to increase hydrocarbon yields in microalgal biofuels production.

CONCLUSION

In this project, the complete transcriptome of the green microalga*, Botryococcus braunii* (Race B, strain Guadeloupe) was sequenced, assembled *de novo* and annotated. To ensure complete coverage of the transcriptome and minimise any effects of circadian and diurnal changes in transcription, the algae were cultured in either a 12 h photoperiod or in constant light and sampled every 4 hours during a 28-hour time-course. The transcriptome was sequenced using the Illumina HiSeq 2000 platform and yielded over 2 billion, paired-end sequence reads of 100 bp in length. Following the *de novo* assembly of 331, 569 transcripts with a mean length of 1, 014 base pairs, the transcriptome was annotated using tblastx search results, assignment of KEGG, GO and EC terms and identification of conserved domains. 273,949 predicted open reading frames generated from the translated transcriptome were annotated with conserved domains.

To ascertain that the transcriptome had sufficient coverage and quality the data were mined to elucidate a previously undescribed molecular pathway in *B. braunii*, but one that had was comprehensively described and well understood in the literature and likely present in the algae, namely the circadial clock gene network. *B.* braunii clock genes were identified by a tblastx search using existing and model clock proteins as queries. Positive BLAST hits were further characterized by identification of conserved domains using Hidden Markov Models to identify conserved, functional domains in the primary amino-acid sequences. This analysis resulted in the discovery of two conserved clock proteins,; comp56012c0seq2 is a homologue of the pseudo response regulators (PRRs) in *Arabidopsis thaliana* and *Ostreococcus tauri* (of which *TOC1* is the archetype), and termed *BbPRR*; comp170985c0seq11 is a homologue of *CCA1/LHY* and was termed *BbCCA1*. A model of the *B. braunii* circadian clock comprising a single transcription/inhibition loop between *BbPRR* and *BbCCA1* was proposed, which is analogous to the *O. tauri* circadian model.

Having confirmed the integrity of the transcriptomic dataset in this manner, differential expression analysis of the annotated *B. braunii* diel transcriptome was performed using raw count data generated by time-series RNA-seq. To ensure minimal identification of false-positives, highly stringent adjusted p value cutoff of 0.01 was used. 287,421 transcripts, *i.e.* 92% of the entire dataset were not differentially expressed over the timecourse and in both light regimes. Comparisons with previous

investigations in green algae suggest that the estimate of 8% differentially expressed genes in *B. braunii* due to photoperiod and the circadian clock is not improbable.

Non-hierarchical clustering of differentially expressed revealed four distinct patterns of diel expression, however, lack of time precluded further analysis.

Mapping the transcriptome to the KEGG database visualised the enzymatic pathways that were predominantly constitutively expressed or differentially expressed in a light:dark (photoperiodic and circadian) or in a constant light (circadian) regime. Some degree of overlap between the diagrams was observed, which was ascribed to the fact that, due to redundancy in the KEGG labeling, multiple transcripts may map to a specific enzyme. Despite this particularity, it was clear that most KEGG pathways combine constitutively and differentially expressed genes. It was concluded that metabolism in *B. braunii*, is parsimoniously controlled by critical enzymatic nodes, rather than by holistic regulation of entire pathways.

B. braunii is most noteworthy by its capacity to synthesize and secrete long-chain, liquid hydrocarbons (botryococcenes) derived from the terpenoid biosynthetic pathway and is unusual in that it partitions carbon in favour of terpenoid and fatty acid synthesis; 45% and 10% of carbon flux, respectively. Using the KEGG database as a reference, the mapped *B. braunii* transcripts reconstructed, fully, MEP/DOXP pathway for botryococcene synthesis, correlating well with previously published biochemical data. However, some constitutively expressed enzymes involved in the Mevalonate pathway were also identified. Surprisingly, it appears that the botryococcene pathway may be differentially regulated throughout the day but it was not possible to clarify the nature of that temporal regulation due to time constraints. Further mining of co-expression data is therefore required to elucidate the precise mechanisms by which this regulation occurs, either by direct gene regulation or indirectly through metabolite partitioning. Hypotheses formulated based on this data can be empirically tested by targeted specific timepoints in the day to extract and analyse hydrocarbon content of *B. braunii*, using GC-MS or Raman Spectroscopy.

Circadian and diurnal controls over metabolic pathways potentially have important implications for the industrial applications of *B. braunii* by allowing the timing of oil extraction and downstream processing methods and to be optimized. Furthermore, the integration of the transcriptomic dataset presented here with data resulting from the metabolic profiling of *B. braunii* could shed light on the regulatory networks represented by the clusters of gene expression observed. As in the case of farnesene, the knowledge of the molecular pathways for hydrocarbon production in *B. braunii* and their regulation may ultimately facilitate the translation of unique *B. braunii*

characteristics into alternative and more tractable algae (or other microbial systems) for the production of sustainable biofuels.

APPENDIX

Table 1 cDNA library primer sequences 224
Figure 1 Bioanalyzer traces of cDNA libraries 0LDP1- 8LDP2 225
Figure 2 Bioanalyzer trace images 8LDP3- 20LDP2 226
Figure 3 Bioanalyzer trace images 20LDP3- 0LLP1 227
Figure 4 Bioanalyzer trace images 0LLP2- 8LLP3 228
Figure 5 Bioanalyzer trace images 12LLP1- 20LLP2 229
Figure 6 Bioanalyzer trace images 20LLP3- 28LLP3 230
Figure 7 Processed read quality 0LDP2- 8LDP2 231
Figure 8 Processed read quality 8LDP3- 20LDP2 232
Figure 9 Processed read quality 20LDP3- 0LLP1 233
Figure 10 Processed read quality 0LLP2- 8LLP3 234
Figure 11 Processed read quality 12LLP1- 20LLP2 235
Figure 12 Processed read quality 20LLP3- 28LLP3 236
Figure 13 Taxonomic distribution of BLAST hits 238
Figure 14 CCA1/ LHY- like clock component HMM alignment 240
Figure 15 PRR1/ TOC1 HMM alignment 241
Figure 16 PRR3 HMM alignment 242
Figure 17 PRR5 HMM alignment 243
Figure 18 PRR7 HMM alignment 244
Figure 19 PRR9 HMM alignment 245
Figure 20 Boxplot of reads overlapping with contigs 246
Table 2 Total reads overlapping contigs 247
Figure 21 Spearman's rank cluster tree of samples 248
Figure 22 Scree plot of LD differentially expressed count data 249
Figure 23 Scree plot of LL differentially expressed count data 249

ScriptSeq primer	Primer sequence (5' to 3')
Forward primer	
1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
2	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
3	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
4	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
5	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
6	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
7	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
8	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
9	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
10	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
11	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
12	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 1 cDNA library primer sequences

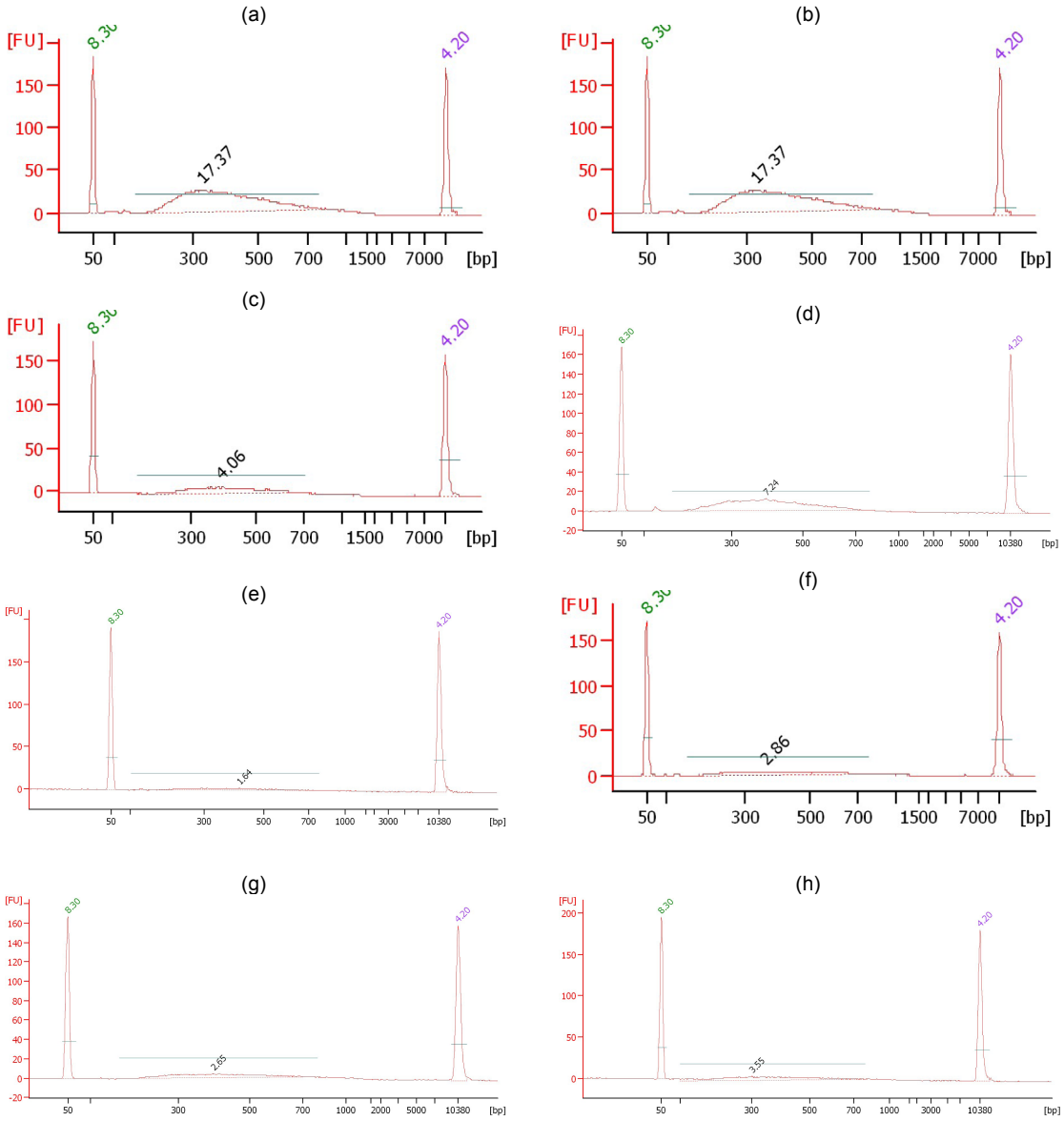


Figure 1 Bioanalyzer traces of cDNA libraries 0LDP1- 8LDP2

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 0LDP1 (a), 0LDP2 (b), 0LDP3 (c), 4LDP1 (d), 4LDP2 (e), 4LDP3 (f), 8LDP1 (g), 8LDP2 (h) are shown. DNA marker peaks are labeled in green and purple, with the 100- 800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

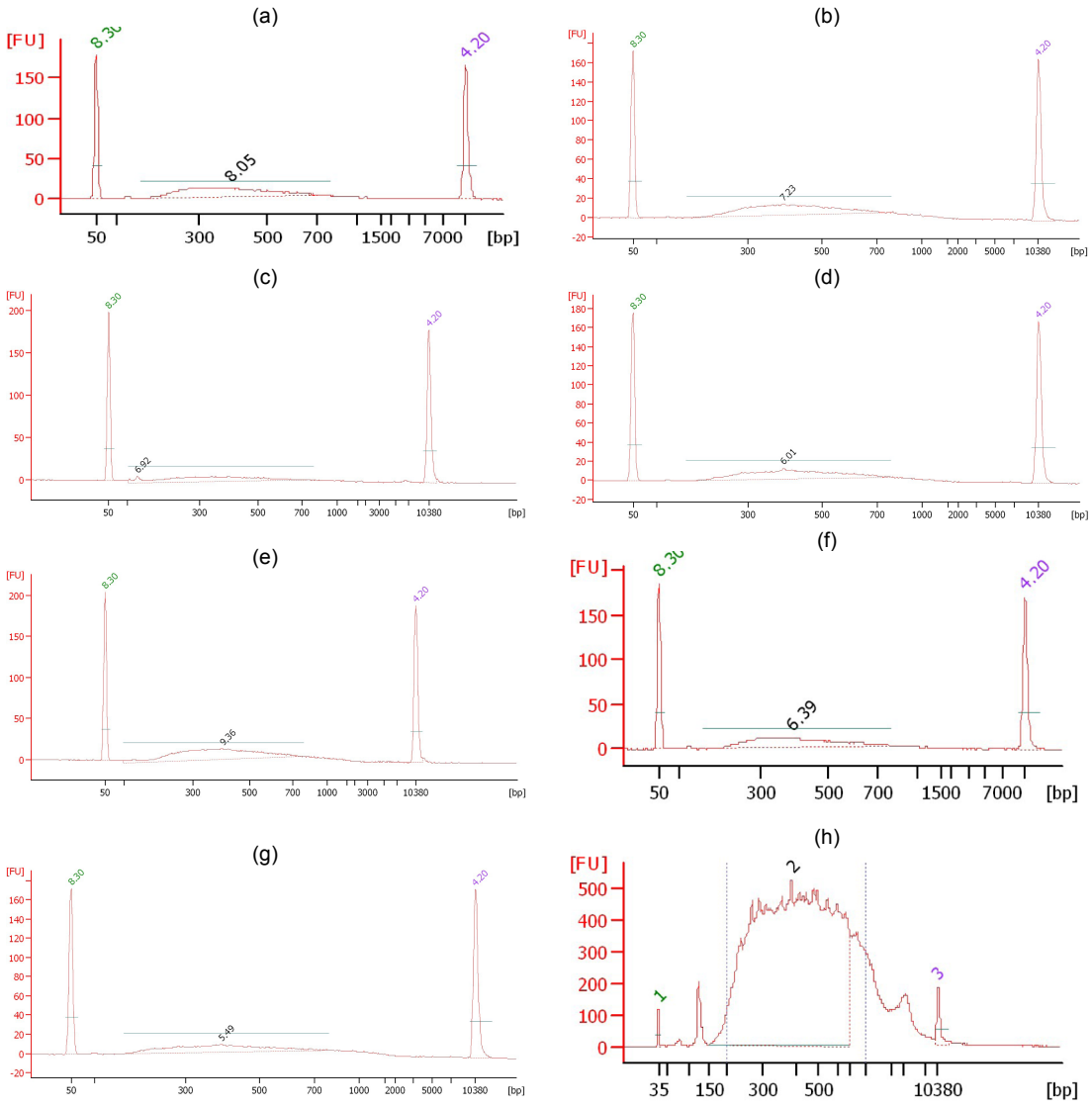


Figure 2 Bioanalyzer trace images 8LDP3- 20LDP2

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 8LDP3 (a), 12LDP1 (b), 12LDP2 (c), 16LDP1 (d), 16LDP2 (e), 16LDP3 (f), 20LDP1 (g) and 20LDP2 (h) are shown. DNA marker peaks are labeled in green and purple, with the 100- 800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

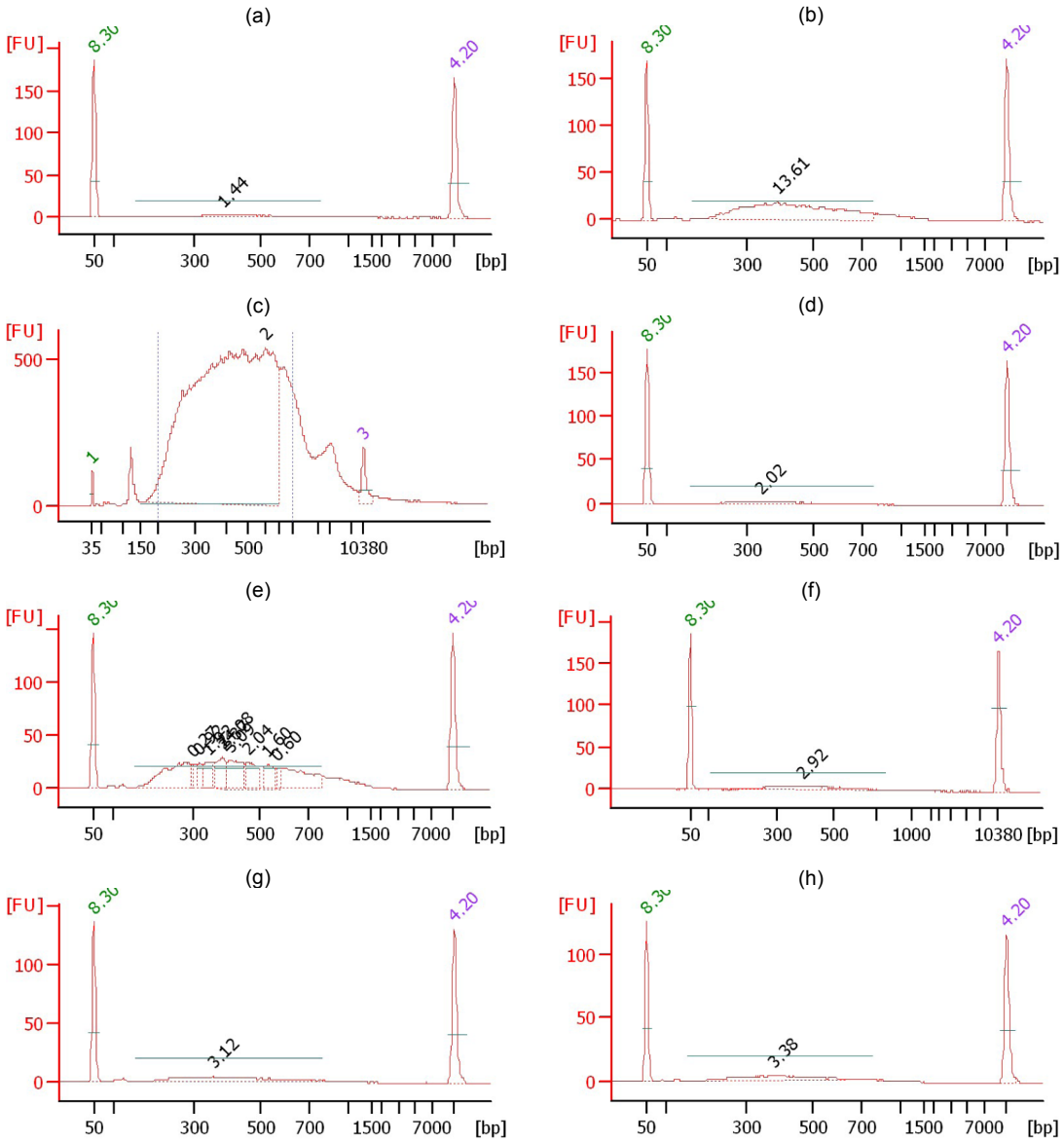


Figure 3 Bioanalyzer trace images 20LDP3- 0LLP1

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 20LDP3 (a), 24LDP1 (b), 24LDP2 (c), 24LDP3 (d), 28LDP1 (e), 28LDP2 (f), 28LDP3 (g) and 0LLP1 (h) are shown. DNA marker peaks are labeled in green and purple, with the 100- 800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

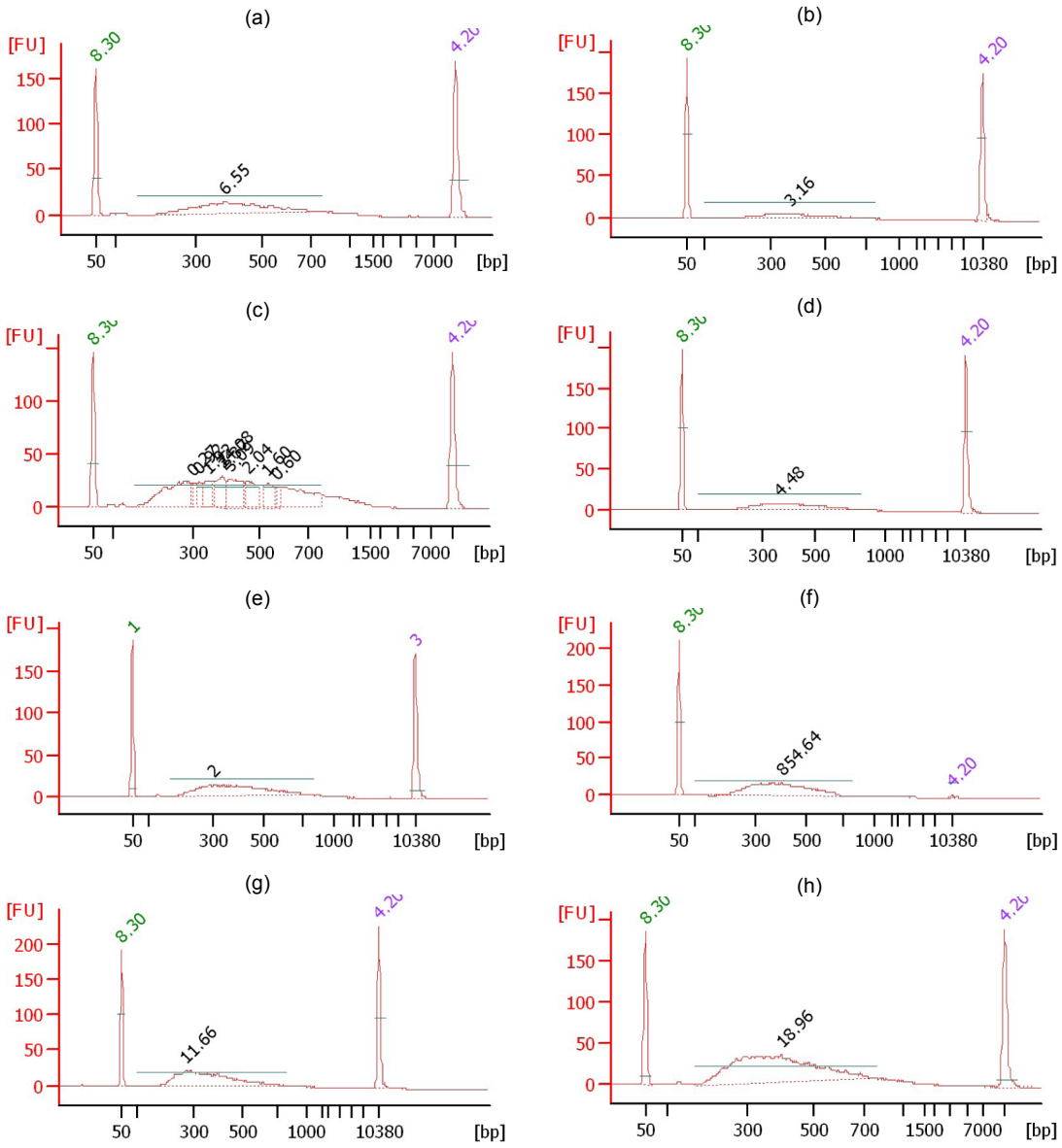


Figure 4 Bioanalyzer trace images 0LLP2- 8LLP3

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 0LLP2 (a), 0LLP3 (b), 4LLP1 (c), 4LLP2 (d), 4LLP3 (e), 8LLP1 (f), 8LLP2 (g), 8LLP3 (h) are shown. DNA marker peaks are labeled in green and purple, with the 100- 800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

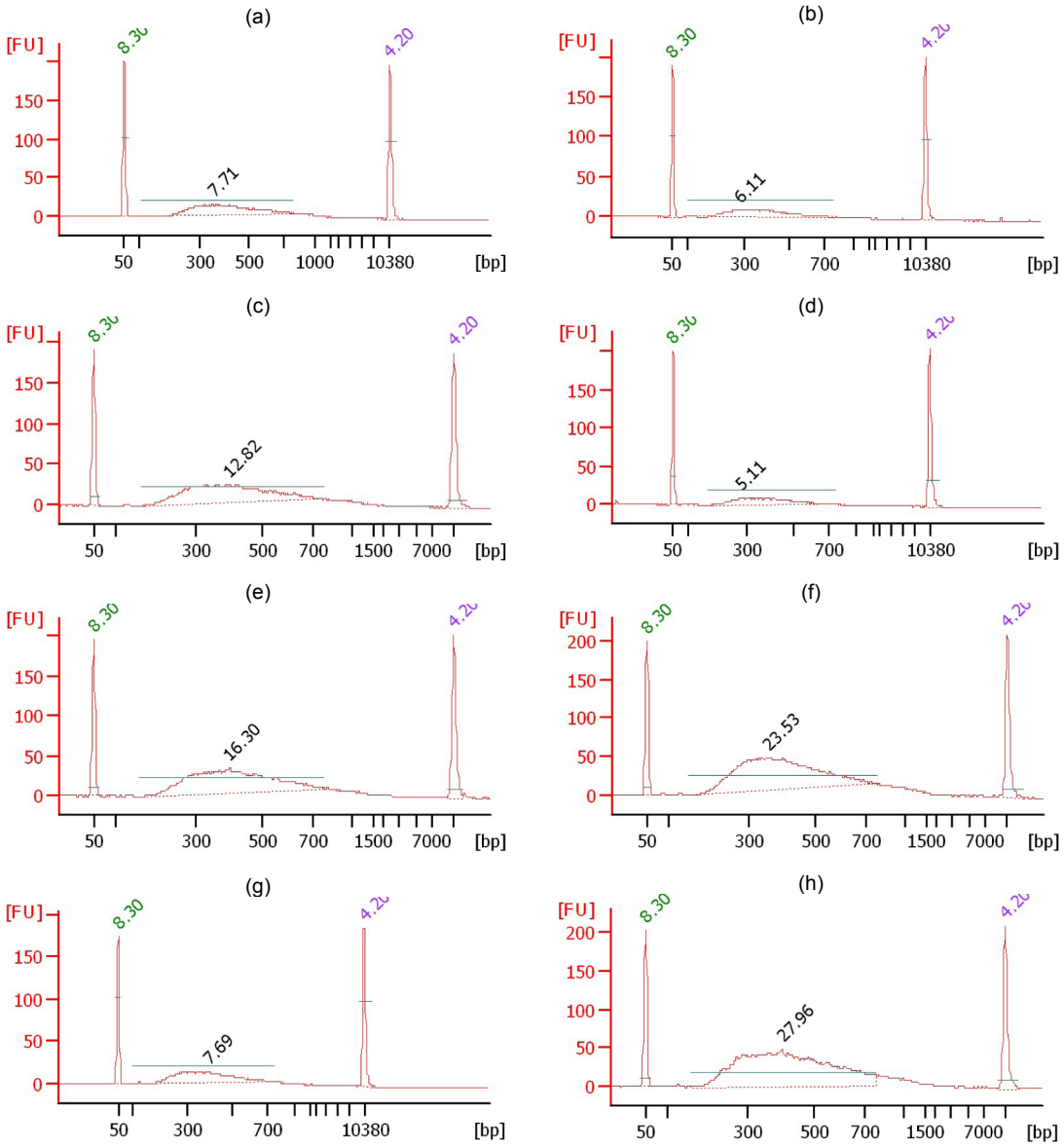


Figure 5 Bioanalyzer trace images 12LLP1- 20LLP2

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 12LLP1 (a), 12LLP2 (b), 12LLP3 (c), 16LLP1 (d), 16LLP2 (e), 16LLP3 (f), 20LLP1 (g), 20LLP2 (h) are shown. DNA marker peaks are labeled in green and purple, with the 100- 800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

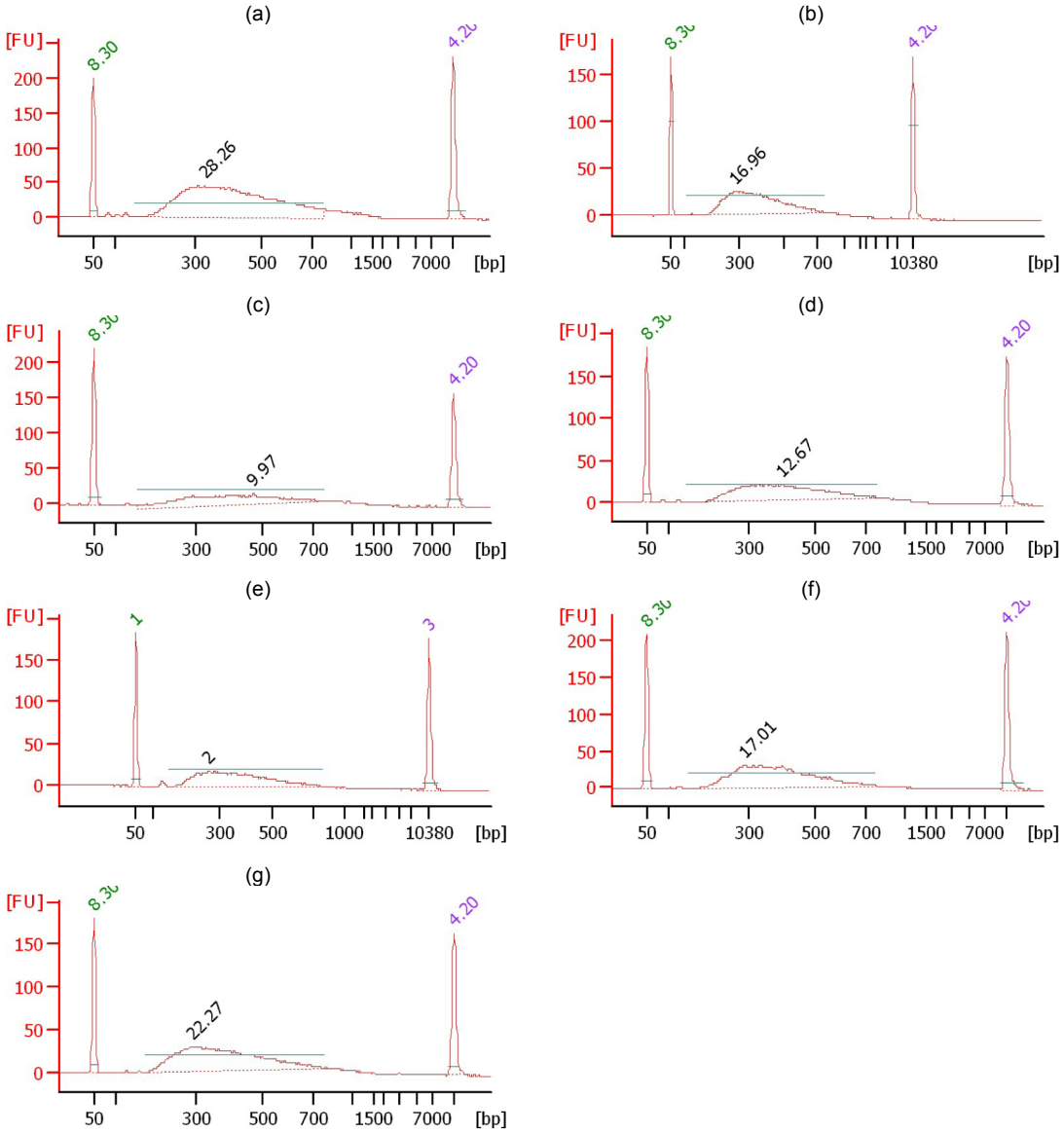


Figure 6 Bioanalyzer trace images 20LLP3- 28LLP3

Plots of fluorescence units (FU) against size in base pairs (bp) for libraries 20LLP3 (a), 24LLP1 (b), 24LLP2 (c), 24LLP3 (d), 28LLP1 (e), 28LLP2 (f), 28LLP3 (g) are shown. DNA marker peaks are labeled in green and purple, with the 100-800 bp cDNA annotated with concentration in ng/μl. Fluorescence units are plotted on the y axis.

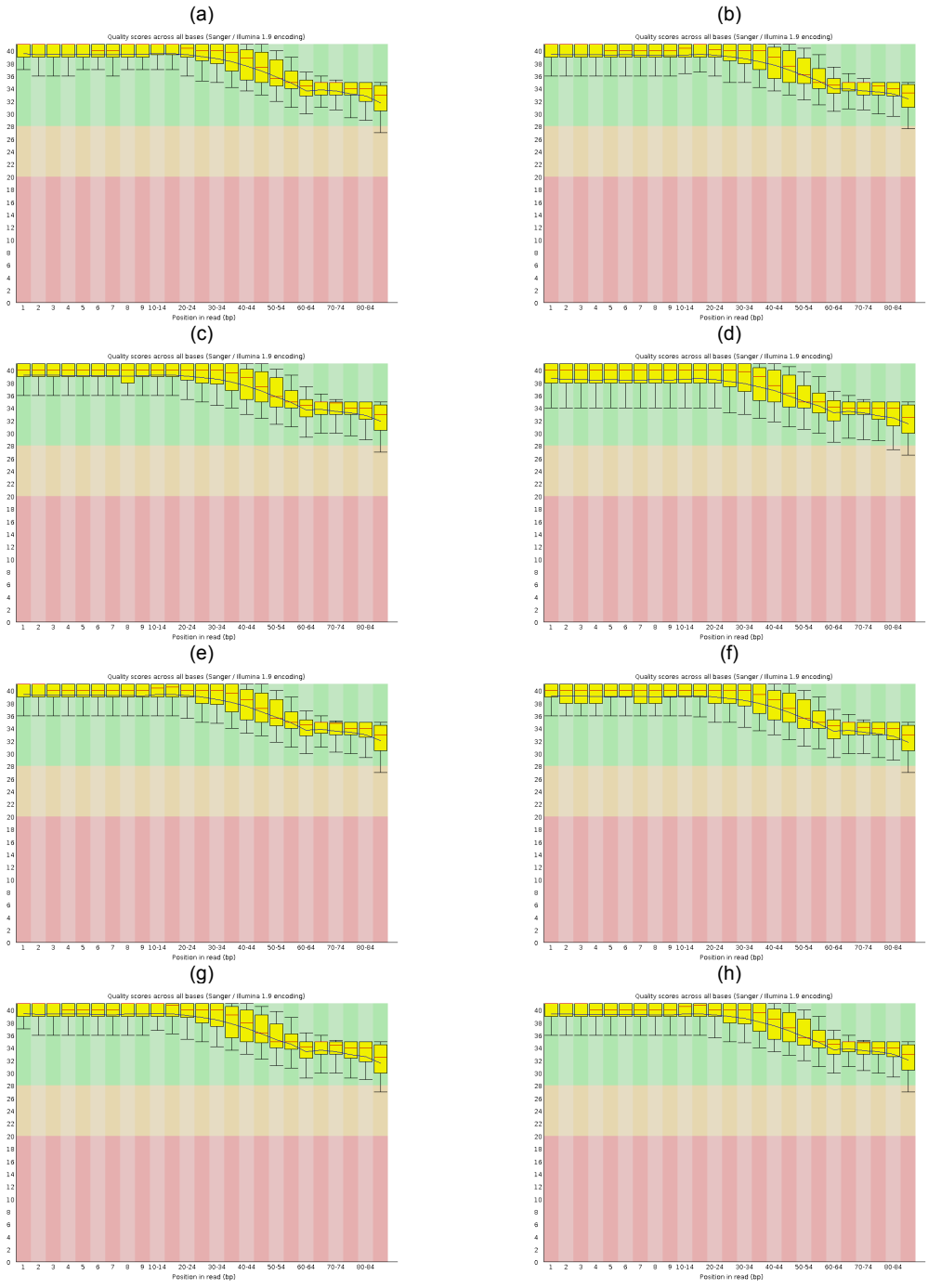


Figure 7 Processed read quality 0LDP2- 8LDP2

FastQC box- whisker plots of per base sequence quality after read processing for samples 0LDP1 (a), 0LDP2 (b), 0LDP3 (c), 4LDP1 (d), 4LDP2 (e), 4LDP3 (f), 8LDP1 (g), 8LDP2 (h). Read 1 is used as a representative sequence of each sample as there was no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

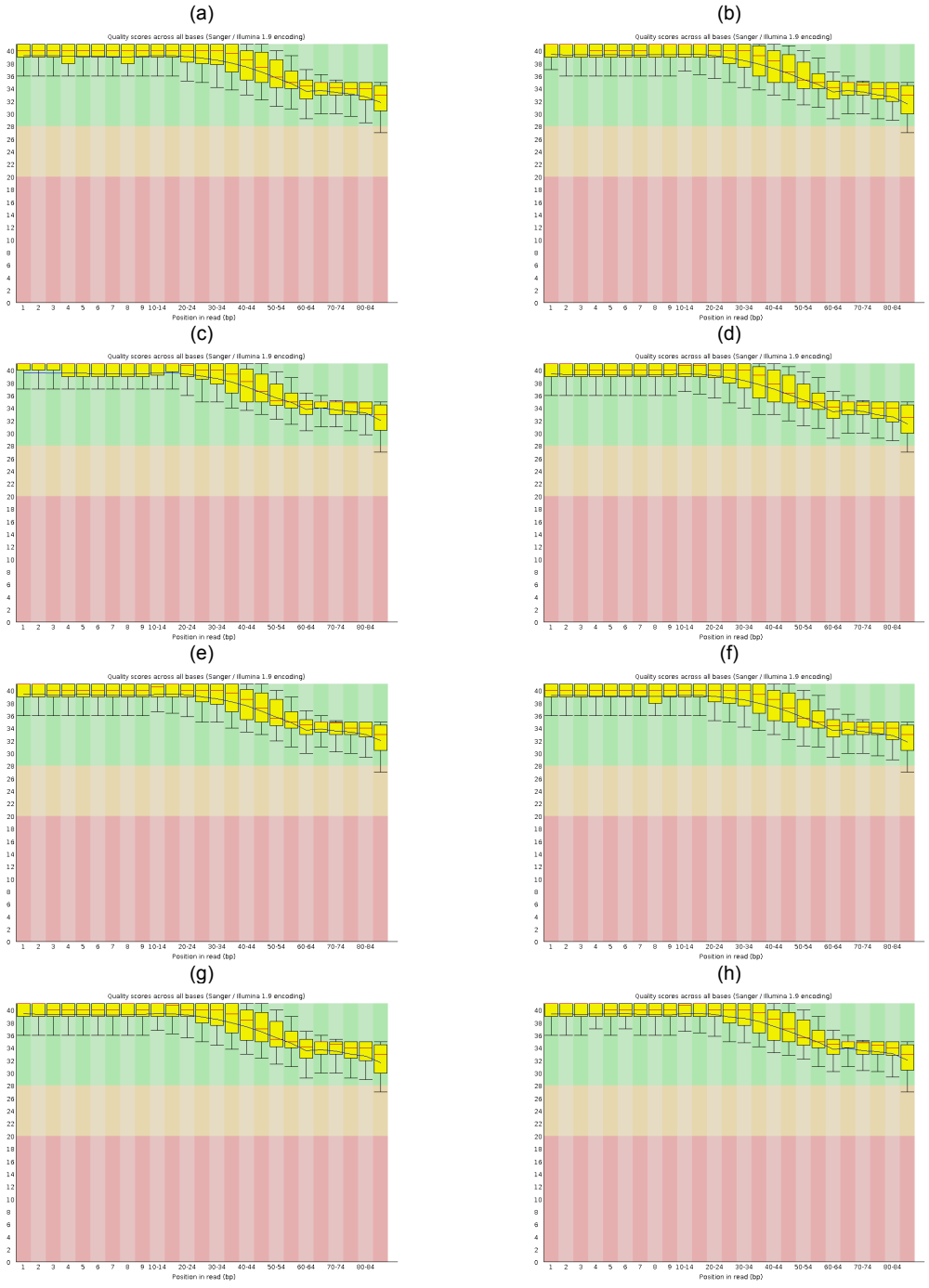


Figure 8 Processed read quality 8LDP3- 20LDP2

FastQC box- whisker plots of per base sequence quality after read processing for samples 8LDP3 (a), 12LDP1 (b), 12LDP2 (c), 16LDP1 (d), 16LDP2 (e), 16LDP3 (f), 20LDP1 (g) and 20LDP2 (h). Read 1 is used as a representative sequence of each sample as there was no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

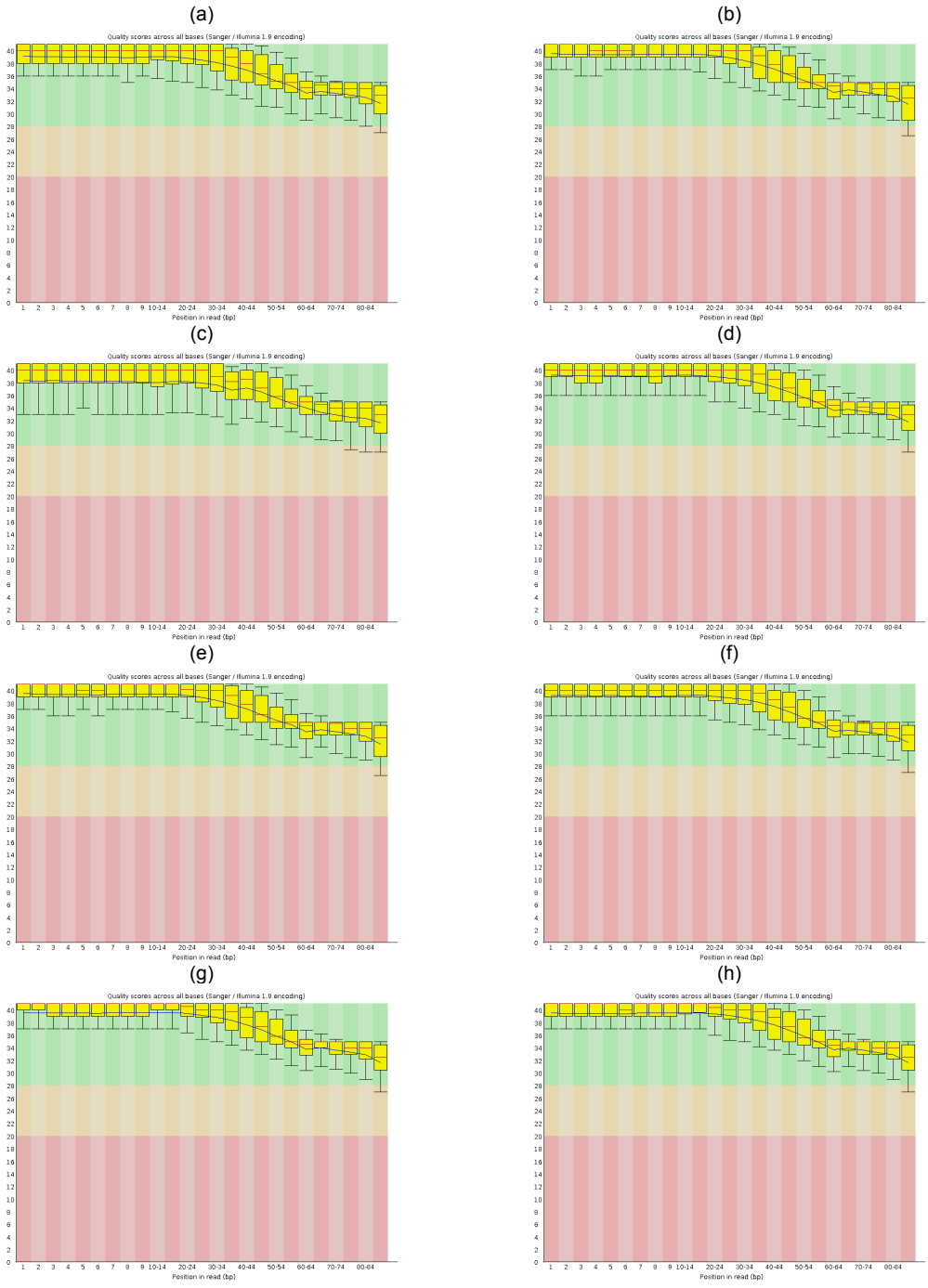


Figure 9 Processed read quality 20LDP3- 0LLP1

FastQC box- whisker plots of per base sequence quality after read processing for samples 20LDP3 (a), 24LDP1 (b), 24LDP2 (c), 24LDP3 (d), 28LDP1 (e), 28LDP2 (f), 28LDP3 (g) and 0LLP1 (h). Read 1 is used as a representative sequence of each sample as there was no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

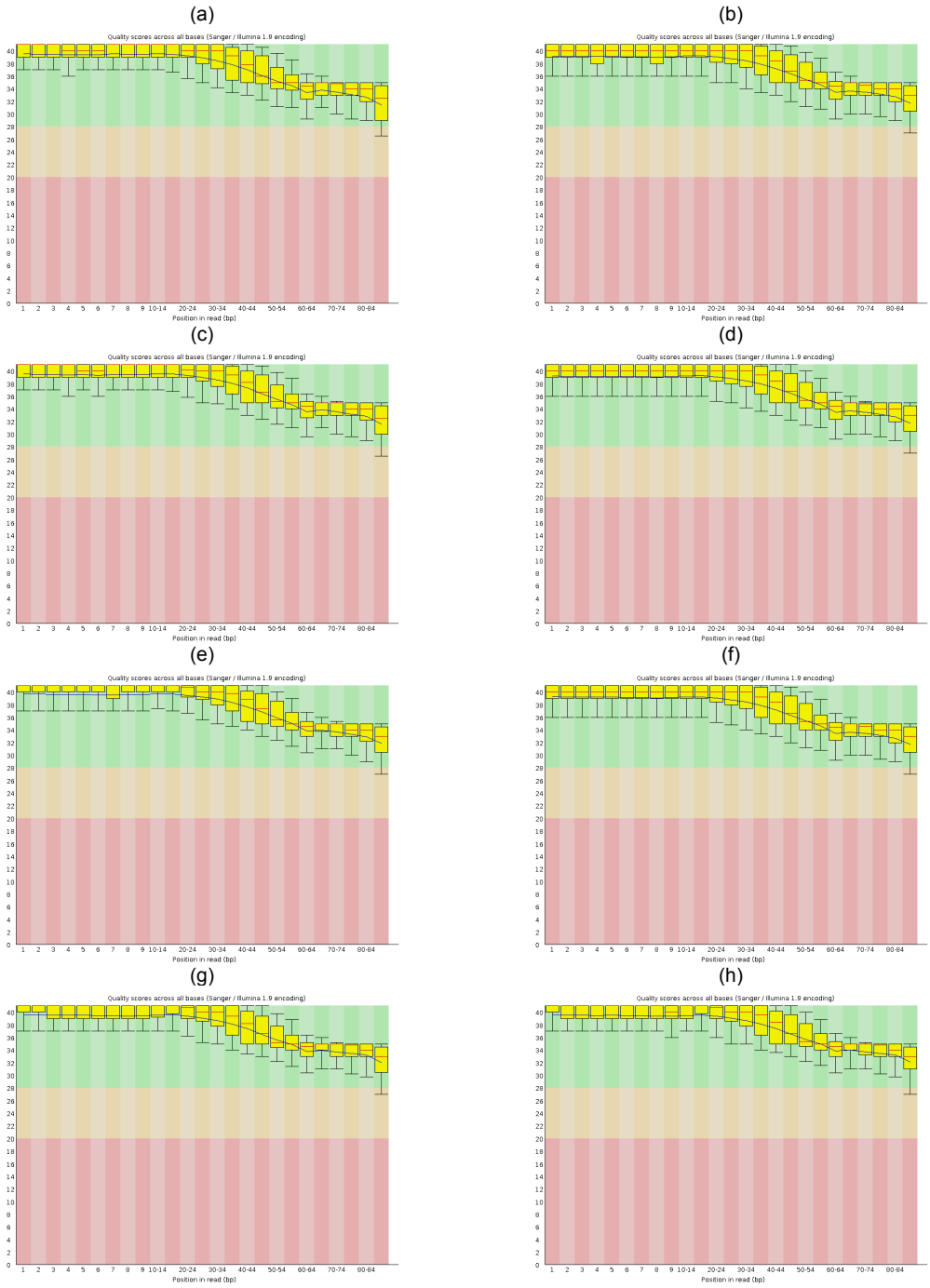


Figure 10 Processed read quality 0LLP2- 8LLP3

FastQC box- whisker plots of per base sequence quality after read processing for samples 0LLP2 (a), 0LLP3 (b), 4LLP1 (c), 4LLP2 (d), 4LLP3 (e), 8LLP1 (f), 8LLP2 (g), 8LLP3 (h). Read 1 is used as a representative sequence of each sample as there was no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

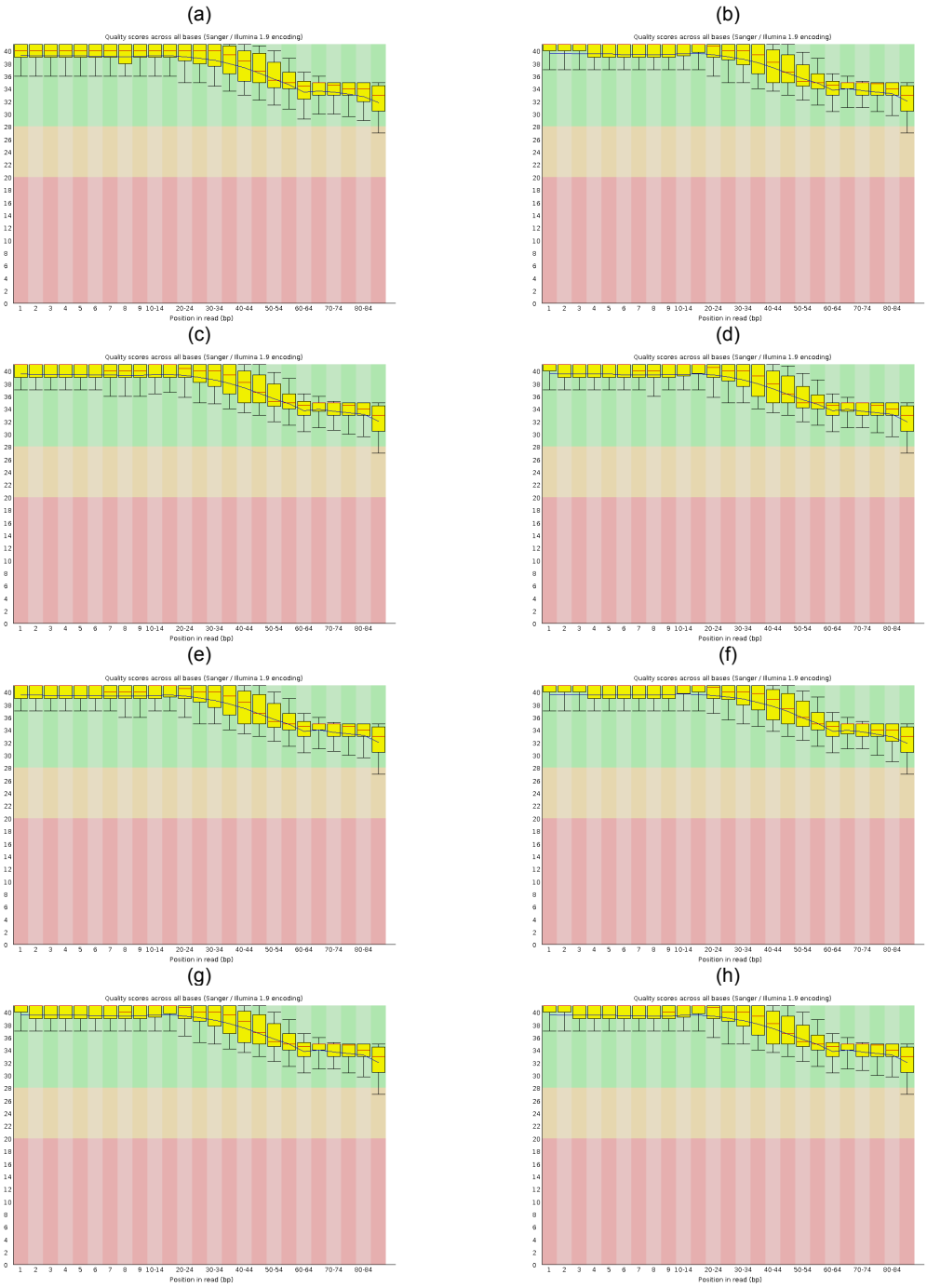


Figure 11 Processed read quality 12LLP1- 20LLP2

FastQC boxplots of per base sequence quality after read processing for samples 12LLP1 (a), 12LLP2 (b), 12LLP3 (c), 16LLP1 (d), 16LLP2 (e), 16LLP3 (f), 20LLP1 (g), 20LLP2 (h). Read 1 is used as a representative sequence of each sample as there were no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

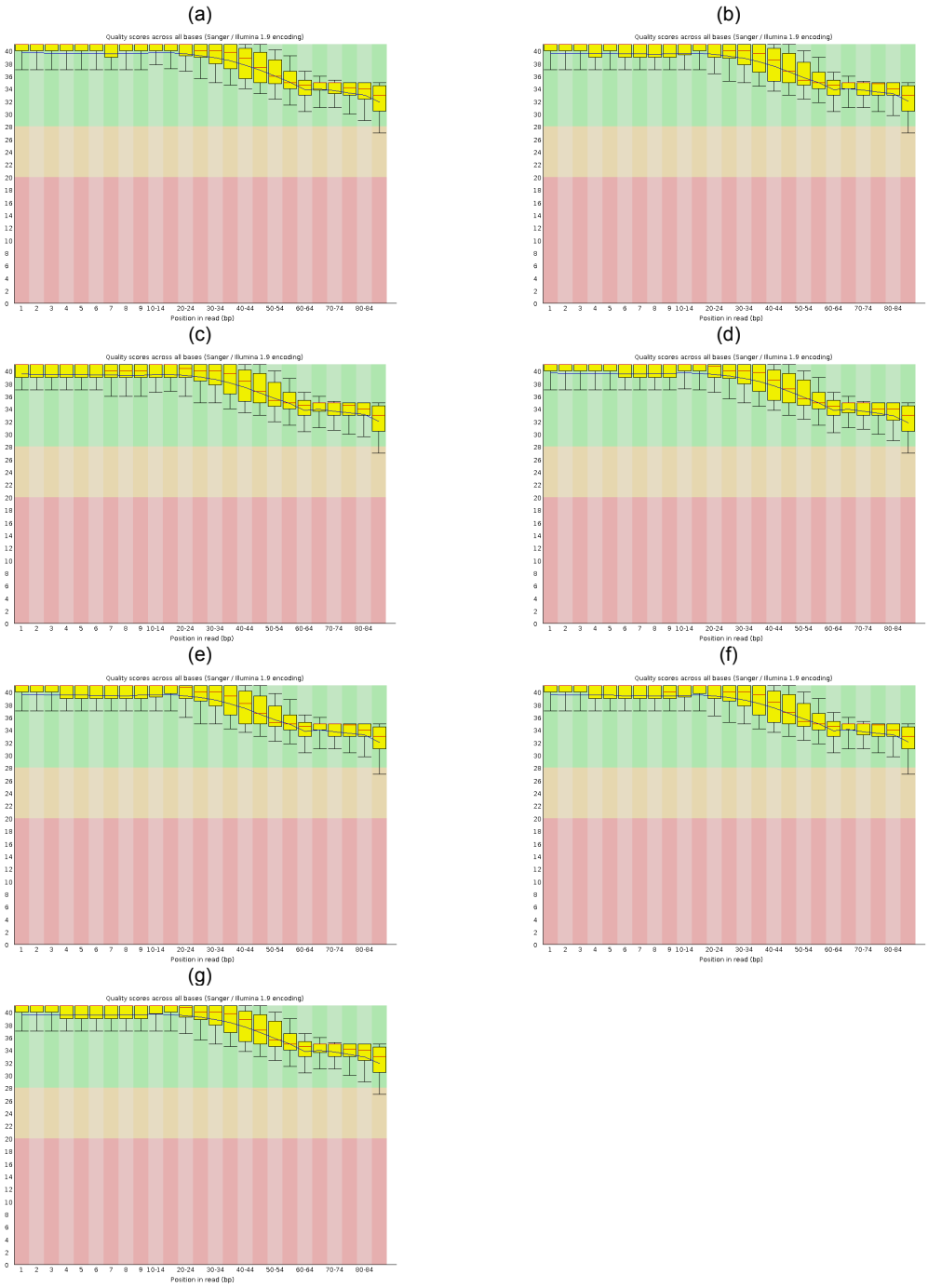


Figure 12 Processed read quality 20LLP3- 28LLP3

FastQC box- whisker plots of per base sequence quality after read processing for samples 20LLP3 (a), 24LLP1 (b), 24LLP2 (c), 24LLP3 (d), 28LLP1 (e), 28LLP2 (f), 28LLP3 (g). Read 1 is used as a representative sequence of each sample as there was no obvious differences between read 1 and 2 and the orphan quality reports. Mean quality score is indicated in blue, median value in red and the interquartile range in yellow with the 10% and 90% values represented by the upper and lower range bar indicators.

Reads total: 331570
Assignments to nodes:
root: 331570
  cellular organisms: 123177
    Bacteria: 20577
      Bacteroidetes/Chlorobi group: 9203
        Bacteroidetes: 9140
          Cytophagia: 3554
            Cytophagales: 3554
              Cytophagaceae: 3234
                Runella: 2049
          Flavobacteriia: 658
            Flavobacteriales: 647
              Flavobacteriaceae: 527
          Sphingobacteriia: 3137
            Sphingobacteriales: 3135
              Chitinophagaceae: 2509
      Chlamydiae/Verrucomicrobia group: 1882
        Verrucomicrobia: 1744
          Verrucomicrobiae: 1212
            Verrucomicrobiales: 1212
              Verrucomicrobiaceae: 1160
                Verrucomicrobium: 1069
      Cyanobacteria: 625
      Firmicutes: 1994
        Clostridia: 1689
          Clostridiales: 1673
      Proteobacteria: 2998
        Alphaproteobacteria: 535
        Gammaproteobacteria: 1248
    Eukaryota: 99827
      Amoebozoa: 7527
        Centramoebida: 2978
          Acanthamoebidae: 2978
            Acanthamoeba: 2978
        Mycetozoa: 2935
          Dictyosteliida: 2895
            Dictyostelium: 1145
      Opisthokonta: 46943
        Fungi: 22145
          Dikarya: 21848
            Ascomycota: 21335
              saccharomyceta: 21273
                Pezizomycotina: 2555
                  leotiomyceta: 2502
                    sordariomyceta: 1907
                      Sordariomycetes: 1858
                        Sordariomycetidae: 1627
                          Magnaporthales: 1468
                            Magnaporthaceae: 1468
                              Magnaporthe: 1310
                Saccharomycotina: 18677
                  Saccharomycetes: 18677
                    Saccharomycetales: 18677
                      mitosporic Saccharomycetales: 5510
                        Candida: 5510
                      Saccharomycetaceae: 13064
                        Nakaseomyces: 11838
        Metazoa: 23179
          Eumetazoa: 22815
            Bilateria: 22381
              Deuterostomia: 2243
                Chordata: 2035
                  Craniata: 1872
                    Vertebrata: 1872
                      Gnathostomata: 1868
                        Teleostomi: 1857
                          Euteleostomi: 1857
                            Sarcopterygii: 1214
                              Tetrapoda: 1214
                                Amniota: 1089
                                  Mammalia: 672
                                    Theria: 663
                                      Eutheria: 613
              Protostomia: 19243
                Ecdysozoa: 19007

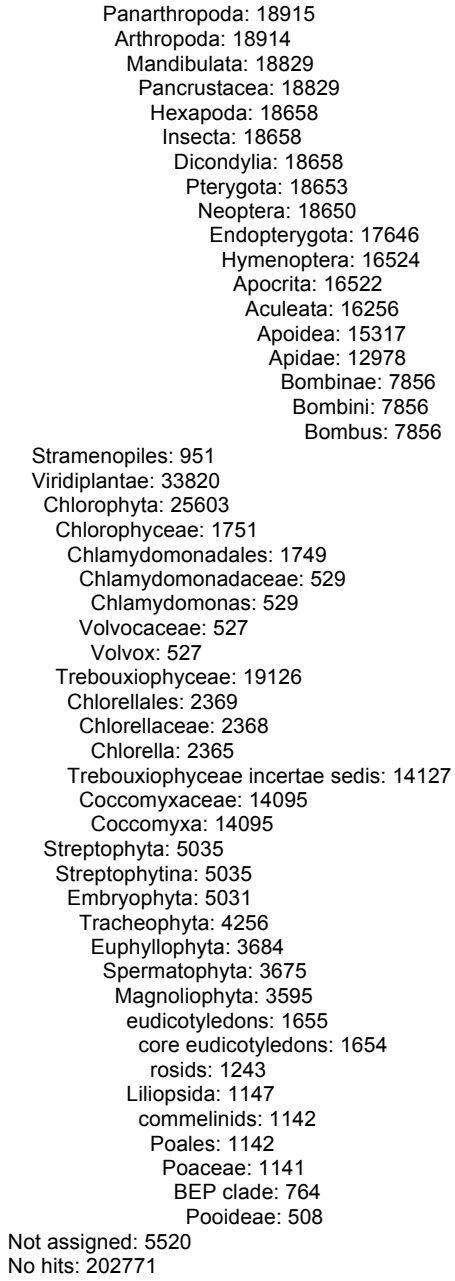
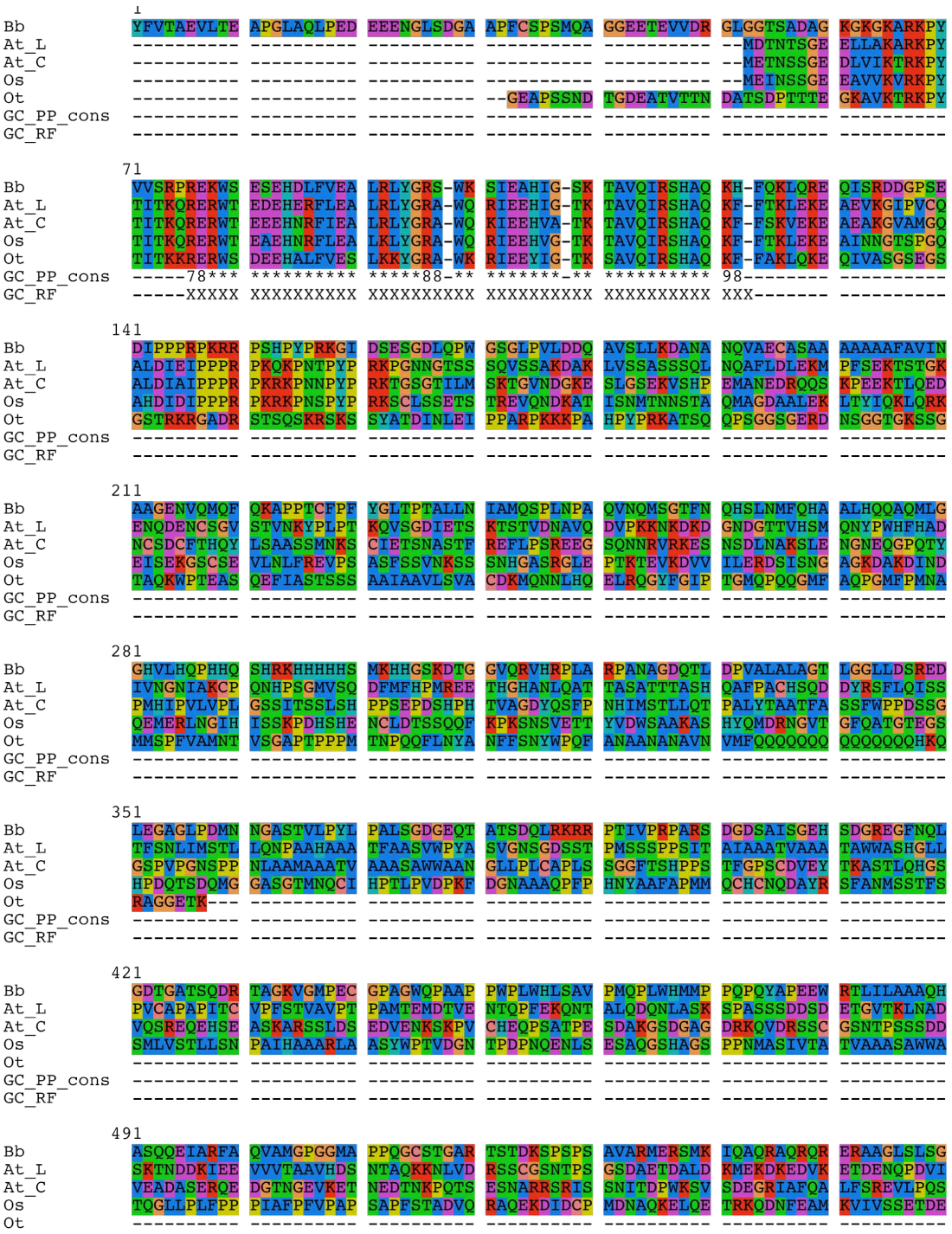


Figure 13 Taxonomic distribution of BLAST hits

The MEGAN summarized number of contigs with BLAST annotations in each taxa. The annotations were filtered to show only taxa supported by 100 contigs or more.



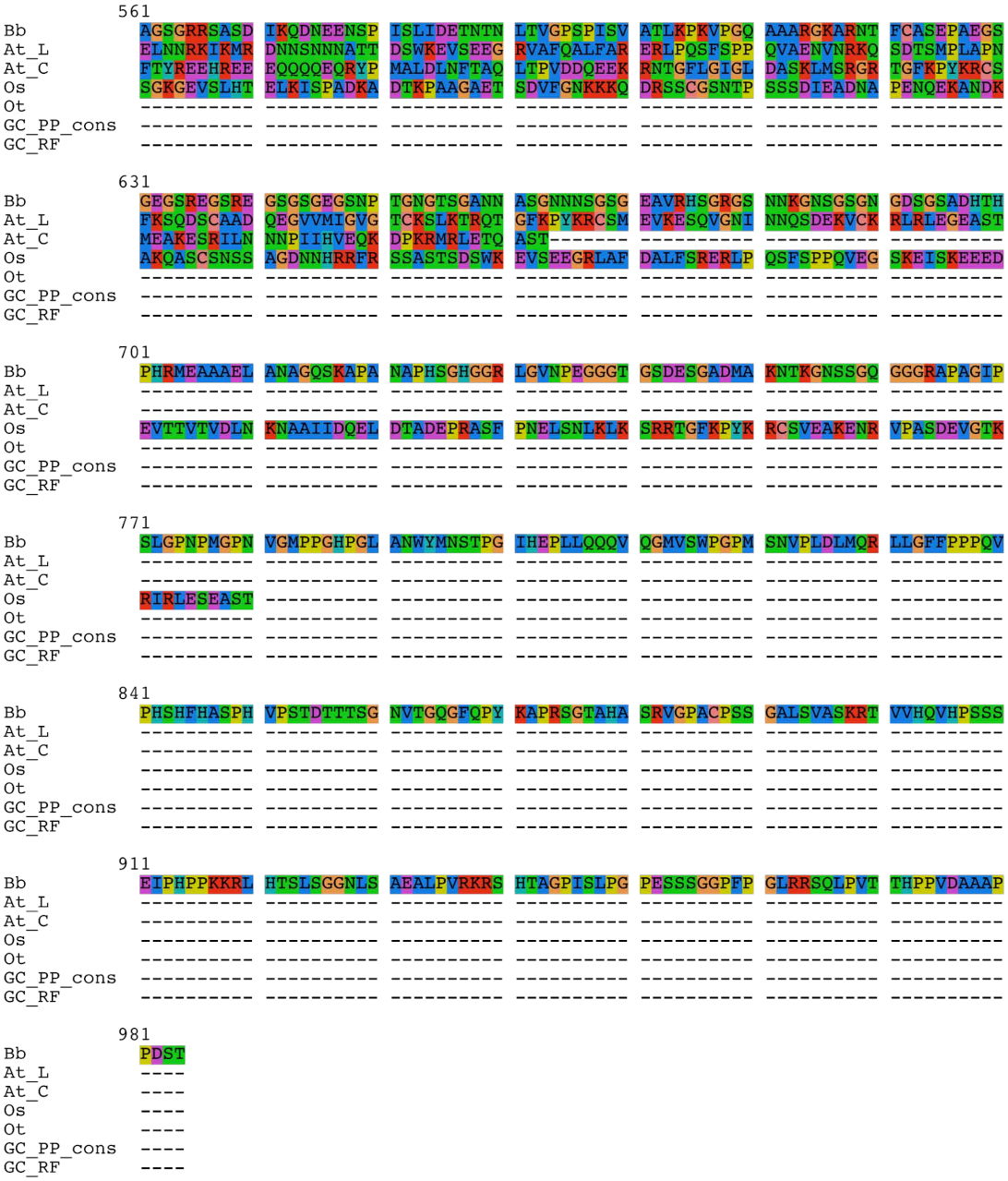


Figure 14 CCA1/ LHY- like clock component HMM alignment (2 pages)

Full- length alignments of the Myb DNA binding domain of LHY/ CCA1 type clock components from *B. braunii*, *A. thaliana*, *O. sativa and O. tauri*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment

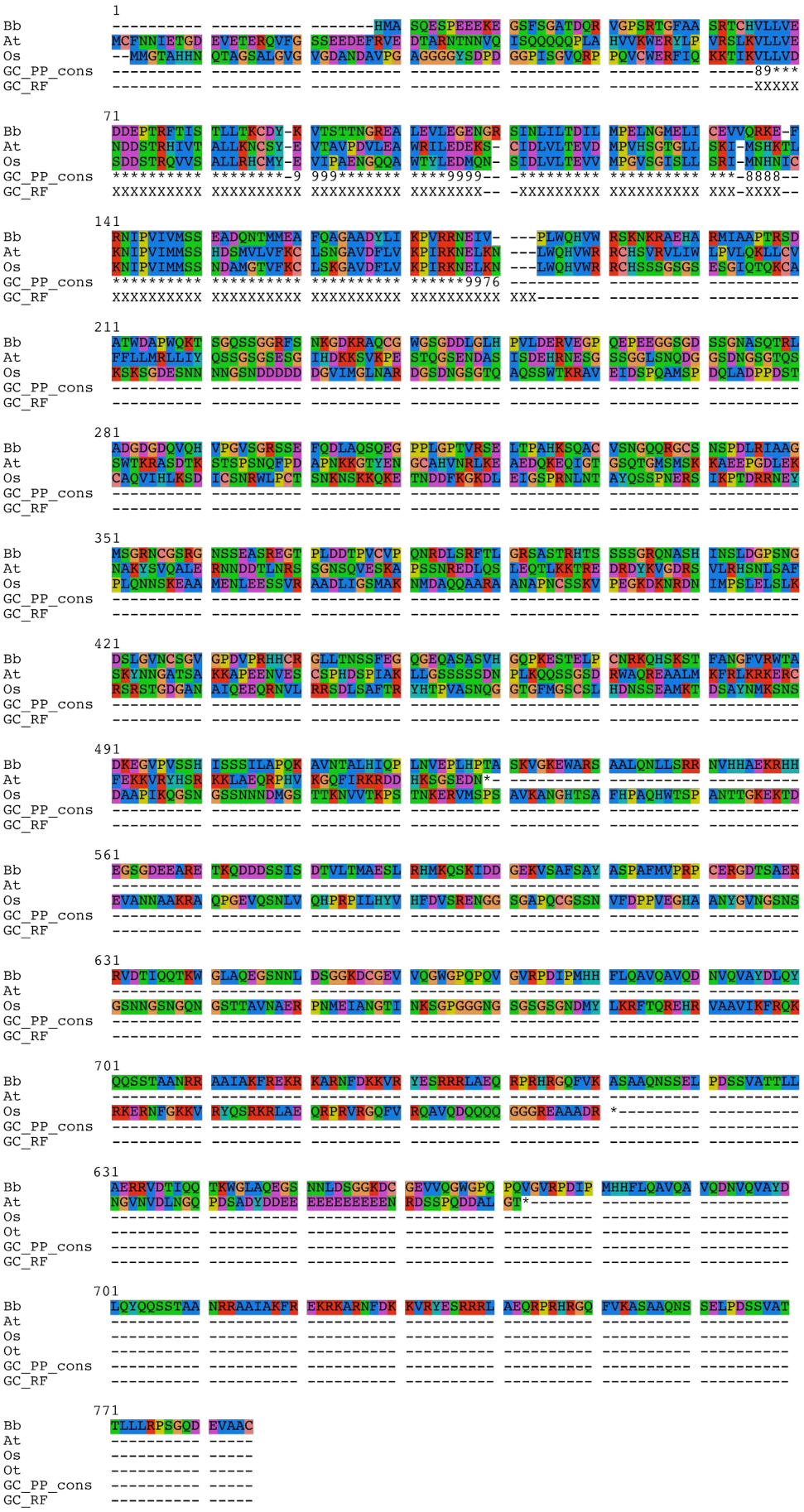


Figure 15 PRR1/ TOC1 HMM alignment

Full- length alignments of the RRR domain of PRR1s (and TOC1) from *B. braunii*, *A. thaliana*, *O. sativa and O. tauri*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment.

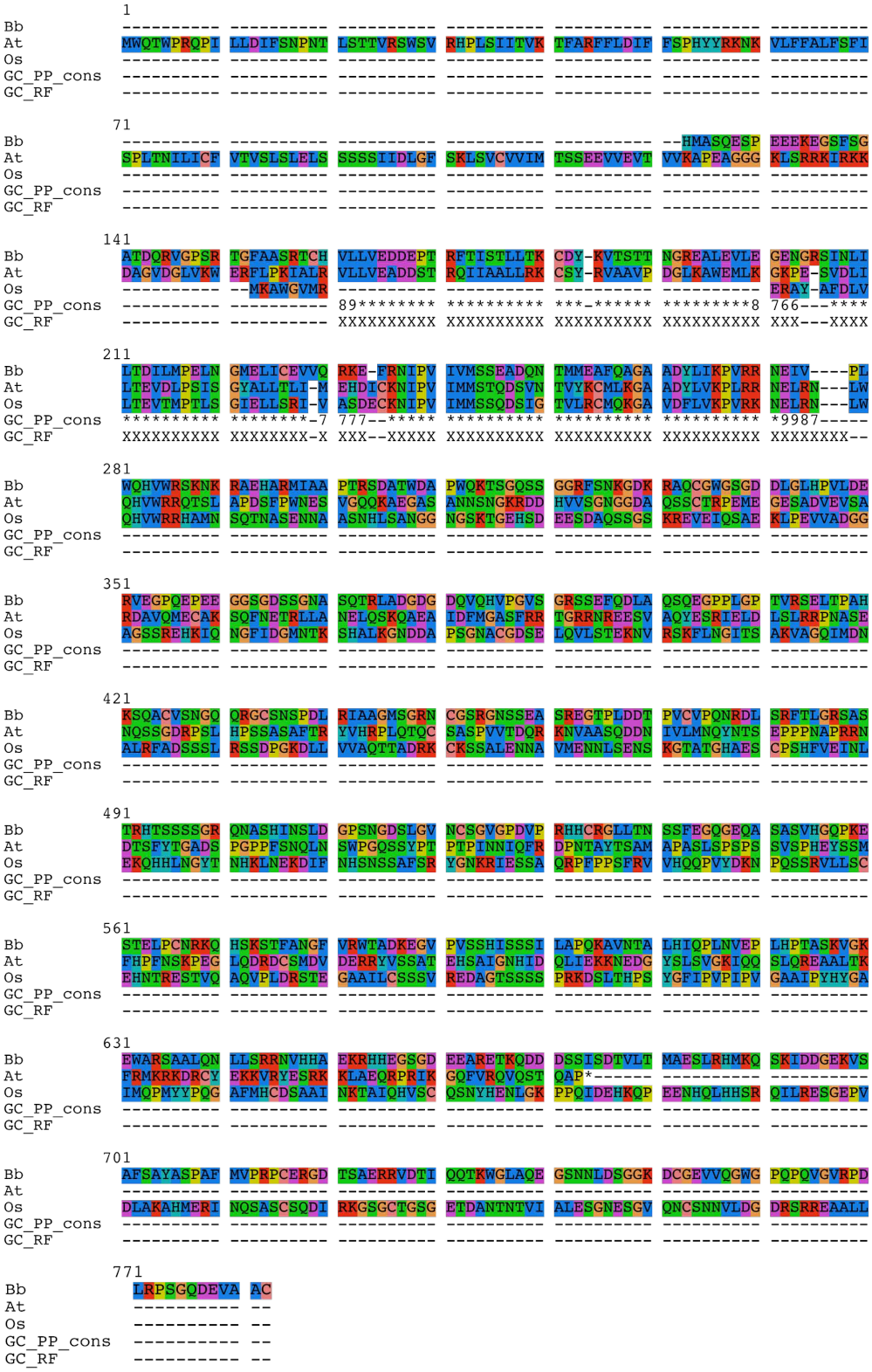


Figure 16 PRR3 HMM alignment

Full- length alignments of the RRR domain of PRR3s from *B. braunii*, *A. thaliana*, and *O. sativa*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment.

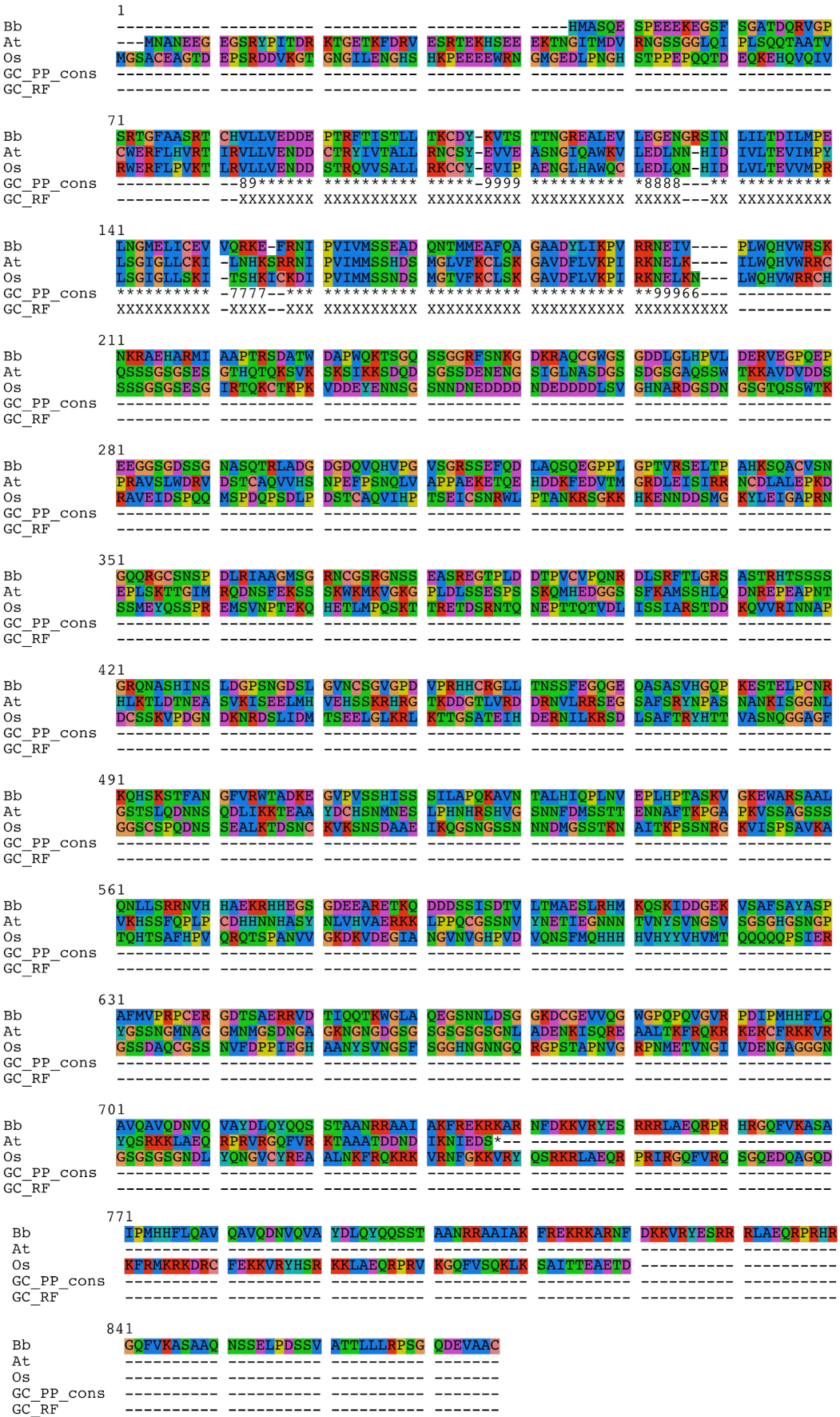


Figure 17 PRR5 HMM alignment

Full- length alignment of the RRR domain of PRR5s from *B. braunii*, *A. thaliana*, and *O. sativa*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment.

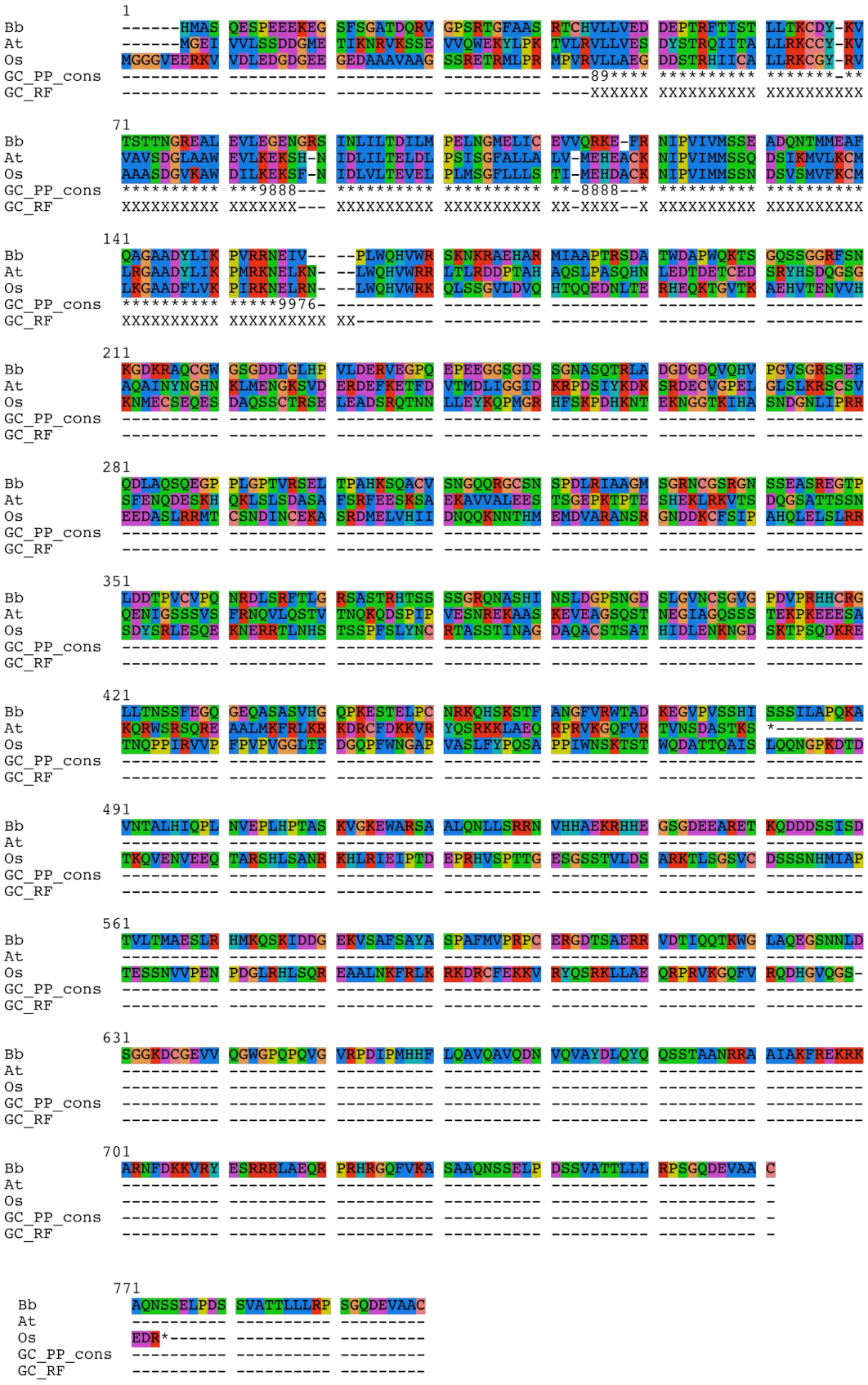


Figure 18 PRR7 HMM alignment

Full- length alignments of the RRR domain of PRR7s from *B. braunii*, *A. thaliana*, and *O. sativa*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment.

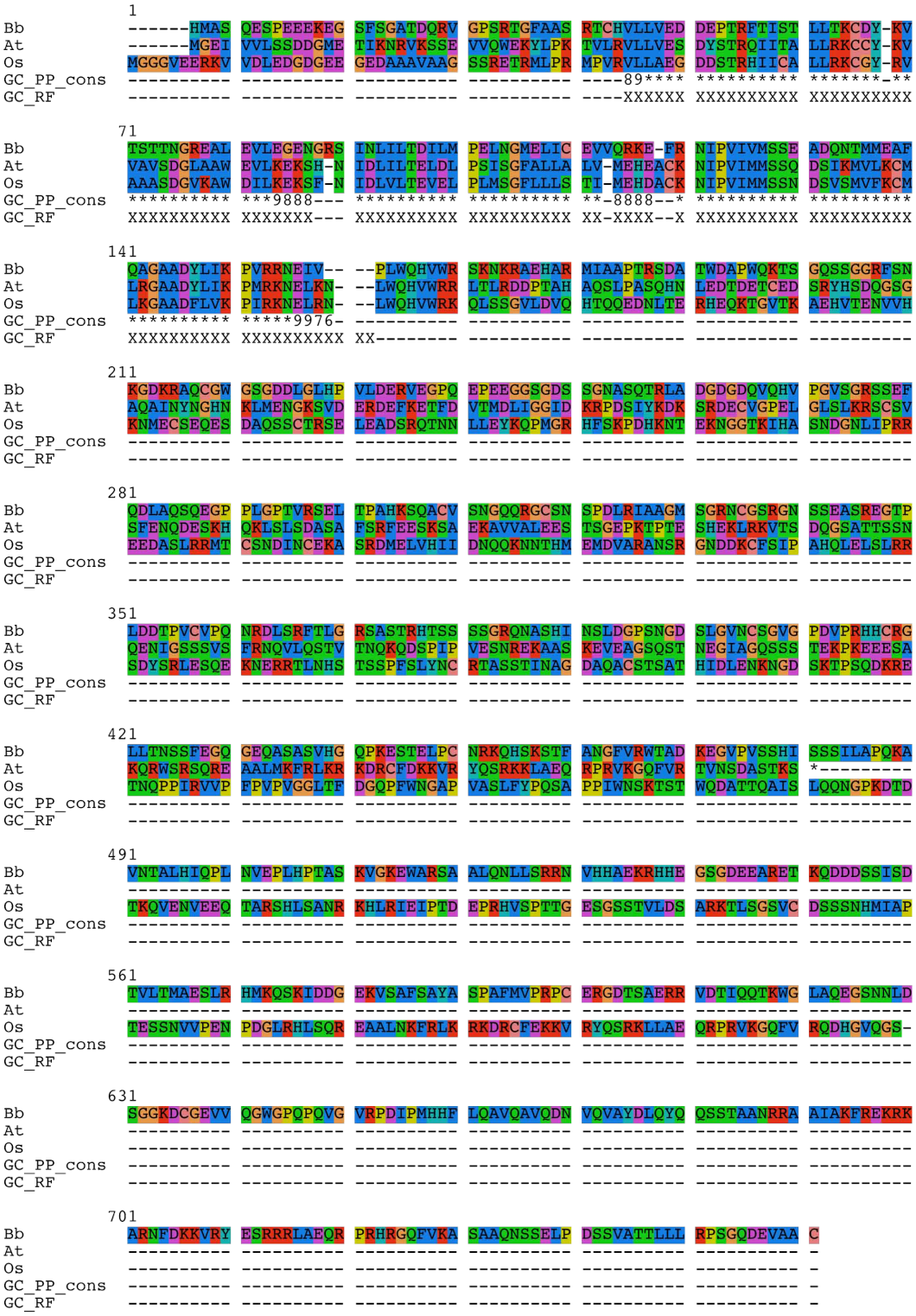


Figure 19 PRR9 HMM alignment

Full- length alignments of the RRR domain of PRR9s from *B. braunii*, *A. thaliana*, and *O. sativa*. The region within the HMM is indicated by an X on the bottom line of each row. The third line is the consensus posterior probability score of alignment.

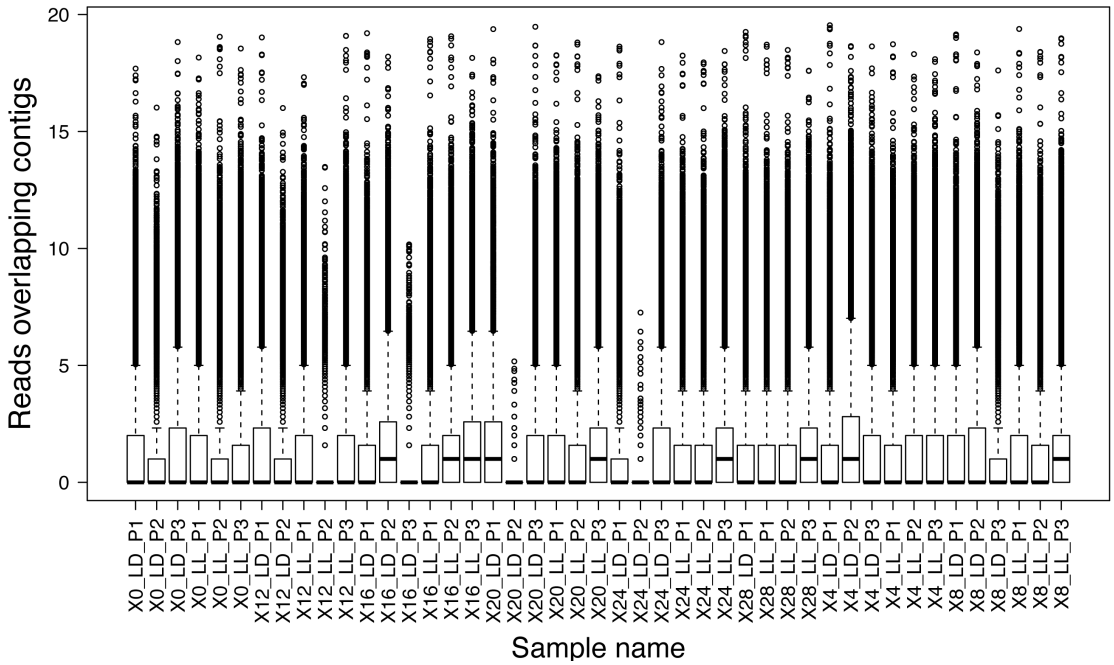


Figure 20 Boxplot of reads overlapping with contigs

The bottom and top of each box represents the first and third quartile of the number of reads counted overlapping with contigs in the *B. braunii* transcriptome per sample. The median value is represented by a black band. Interquartile range (IQR) is indicated by a dashed line and read numbers outside of the range given by 1.5(IQR) are plotted as circles. Sample outliers, 12LLP2, 16LDP3, 20LDP2 and 24LDP2 are clearly seen with low aligned reads.

Sample	Number of reads mapped to reference assembly
0LDP1	8,278,778
0LDP3	13,581,671
0LLP1	9,327,970
0LLP2	5,818,076
0LLP3	8,143,824
4LDP1	7,655,298
4LDP2	17,215,937
4LDP3	8,362,976
4LLP1	6,603,547
4LLP2	7,494,070
4LLP3	10,044,466
8LDP1	7,400,989
8LDP2	10,199,158
8LDP3	3,348,143
8LLP1	8,762,824
8LLP2	5,737,712
8LLP3	9,136,670
12LDP1	7,248,100
12LDP2	2,012,710
12LLP1	6,428,510
12LLP2	315,982
12LLP3	9,108,057
16LDP1	5,520,812
16LDP2	10,888,145
16LDP3	244,388
16LLP1	6,661,734
16LLP2	8,303,298
16LLP3	11,555,116
20LDP1	8,667,512
20LDP3	7,746,621
20LLP1	8,415,967
20LLP2	7,560,650
20LLP3	8,767,407
24LDP1	4,606,796
24LDP3	8,007,143
24LLP1	6,504,743
24LLP2	6,295,523
24LLP3	11,372,279
28LDP1	8,124,746
28LDP2	8,016,524
28LDP3	8,206,089
28LLP1	5,953,470
28LLP2	6,751,452
28LLP3	9,387,419
Total	339,783,302

Table 2 Total reads overlapping contigs

The total number of reads counted overlapping with contigs in the *B. braunii* transcriptome per sample, given by summed counts from HTSeq-count.

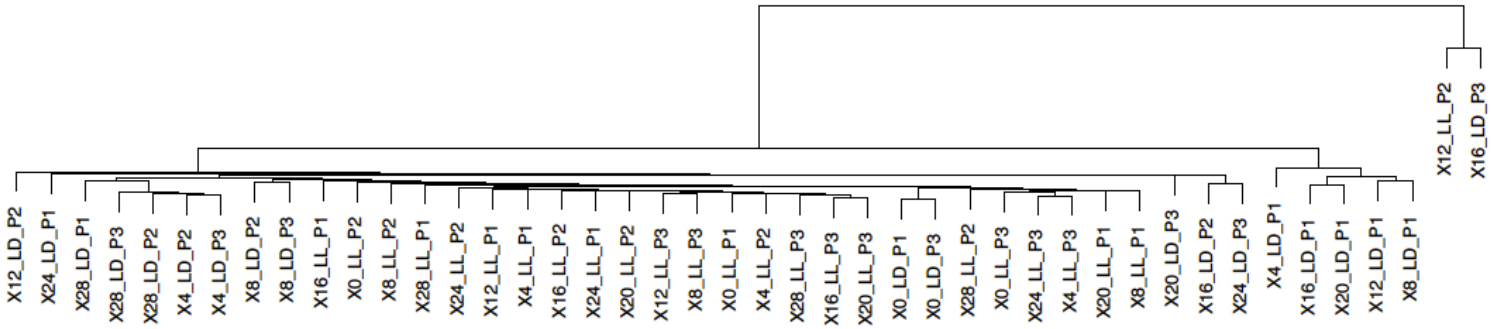


Figure 21 Spearman's rank cluster tree of samples

Spearman's rank correlation and distance metrics (1- correlation value) were used to draw a cluster tree of samples according to their expression profiles. Outliers 12LLP2 and 16LDP3 are clearly shown clustering together, distant from the rest. This analysis was performed after 0LDP2, 20LDP2 and 24LDP2 were removed.

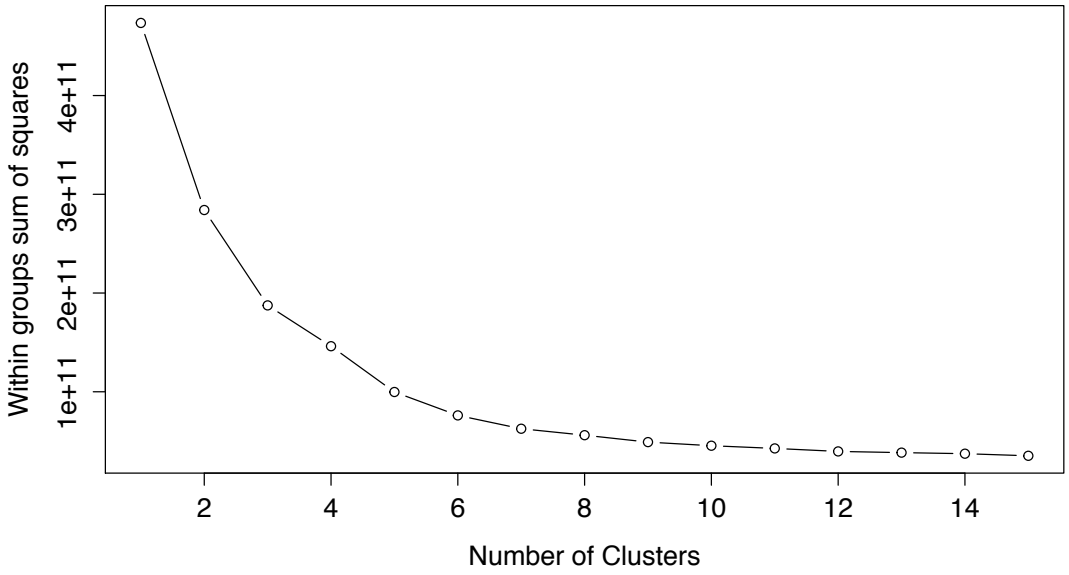


Figure 22 Scree plot of LD differentially expressed count data
Sum of squared variance (Y axis) is plotted for contig expression patterns partitioned into 1 to 15 groups with the closest mean (X axis) for samples under 12:12 photoperiod.

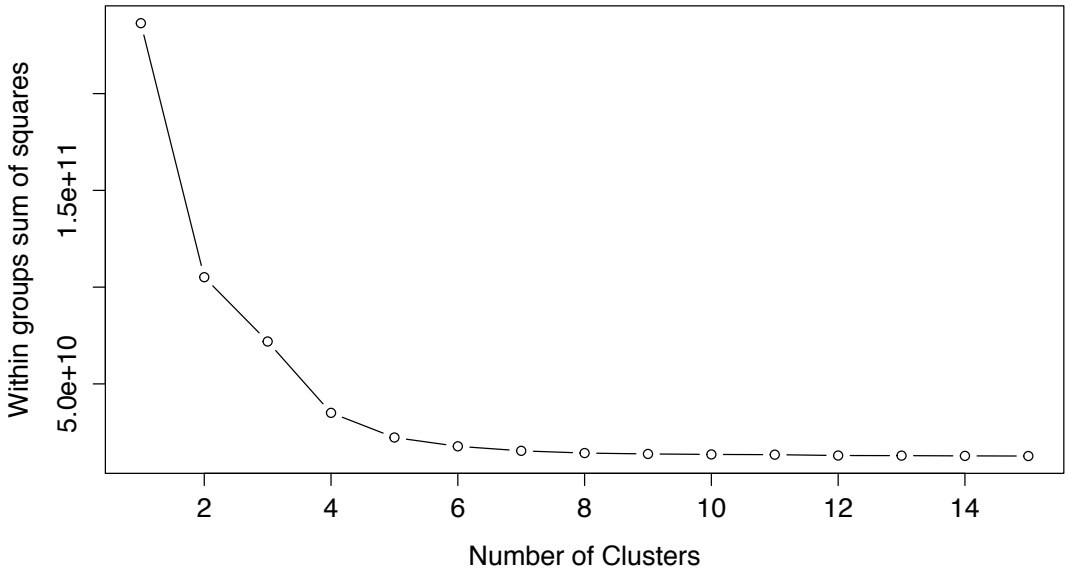


Figure 23 Scree plot of LL differentially expressed count data
Sum of squared variance (Y axis) is plotted for contig expression patterns partitioned into 1 to 15 groups with the closest mean (X axis) for samples under continuous light.

BIBLIOGRAPHY

Achitouv, E., Metzger, P., Rager, M.-N., and Largeau, C. (2004). C31–C34 methylated squalenes from a Bolivian strain of *Botryococcus braunii*. *Phytochemistry* 65: 3159–3165

Alabadi, D. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the *Arabidopsis* circadian clock. *Sscience* 293: 880–883

Arbib, Z., Ruiz, J., Álvarez-Díaz, P., Garrido-Pérez, C., Barragan, J., and Perales, J.A. (2013). Long term outdoor operation of a tubular airlift pilot photobioreactor and a high rate algal pond as tertiary treatment of urban wastewater. *Ecological Engineering* 52: 143–153

Baggs, J.E., Price, T.S., DiTacchio, L., Panda, S., Fitzgerald, G.A., and Hogenesch, J.B. (2009). Network Features of the Mammalian Circadian Clock. *PLoS Biology* 7

Ball, S. G. (2005). Eukaryotic Microalgae Genomics. The Essence of Being a Plant. *Plant Physiology*, 137 (2): 397–398. doi:10.1104/pp.104.900136

Banerjee, A., Sharma, R., Chisti, Y., and Banerjee, U.C. (2002). *Botryococcus braunii*: A renewable source of hydrocarbons and other chemicals. *Critical Reviews in Biotechnology* 22: 245–279

Becker, E.W. (1994). Microalgae: Biotechnology and Microbiology (Cambridge: Cambridge University Press)

Bell-Pederson, D., Cassone, V.M., Earnest, D.J., Golden, S.S., Hardin, P.E., Thomas, T.L., and Zoran, M.J. (2005). Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nature Reviews Genetics* 6: 544–556

Ben-Amotz, A., and Avron, M. (1990). The biotechnology of cultivating the halotolerant alga *Dunaliella*. *Trends in Biotechnology* 8: 121–126

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 25 (22): 3045–3046. doi:10.1093/bioinformatics/btp536

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., *et al*. (2009). *De novo* transcriptome assembly with ABySS. *Bioinformatics* 25: 2872–2877

Bisova, K. (2005). Genome-Wide Annotation and Expression Profiling of Cell Cycle Regulatory Genes in *Chlamydomonas reinhardtii*. *Plant Physiology* 137 (2): 475–491 doi:10.1104/pp.104.05415

Blanco, A.M., Moreno, J., Del Campo, J.A., Rivas, J., and Guerrero, M.G. (2007). Outdoor cultivation of lutein-rich cells of *Muriellopsis sp*. in open ponds. *Applied Microbiol Biotechnol.* 73 (6): 1259-1266

Boothroyd, C.E., Wijnen, H., Naef, F., Saez, L., and Young, M.W. (2007). Integration of Light and Temperature in the Regulation of Circadian Gene Expression in *Drosophila*.

PLoS Genet 3: pp54

Borowitzka, M.A. (1999). Economic evaluation of microalgal processes and products. In Chemicals From Microalgae, Z. Cohen, ed. (London: Taylor & Francis), pp. 387–409
Borowitzka, M.A. (2013). High-value products from microalgae- their development and commercialisation. *Journal of Applied Phycology* 25: 743–756

Cabanelas, I.T.D., Ruiz, J., Arbib, Z., Chinalia, F.A., Garrido-Pérez, C., Rogalla, F., Nascimento, I.A., and Perales, J.A. (2013). Comparing the use of different domestic wastewaters for coupling microalgal production and nutrient removal. *Bioresource Technology* 131: 429–436

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10 (1): 421 doi:10.1186/1471-2105-10-421

Camera, S.L., Gouzerh, G., Dhondt, S., Hoffman, L., Fritig, B., Legrand, M., and Heitz, T. (2004). Metabolic reprogramming in plant innate immunity: the contributions of phenylpropanoid and oxylipin pathways. *Immunological Reviews* 198: 267–284

Campo, J.A., García-González, M., and Guerrero, M.G. (2007). Outdoor cultivation of microalgae for carotenoid production: current state and perspectives. *Applied Microbiology and Biotechnology* 74: 1163–1174

Carbonell-Bejerano, P., Rodriguez, V., Royo, C., Hernáiz, S., Moro-González, L.C., Torres-Viñals, M., and Martínez-Zapater, J.M. (2014). Circadian oscillatory transcriptional programs in grapevine ripening fruits. *BMC Plant Biology* 14: 78–93

Chisti, Y. (2007). Biodiesel from microalgae. *Biotechnology Advances* 25: 294–306

Chiu, S.-Y., Kao, C.-Y., Huang, T.-T., Lin, C.-J., Ong, S.-C., Chen, C.-D., Chang, J.-S., and Lin, C.-S. (2011). Microalgal biomass production and on-site bioremediation of carbon dioxide, nitrogen oxide and sulfur dioxide from flue gas using *Chlorella sp*. cultures. *Bioresource Technology* 102: 9135–9142

Chow, B.Y., and Kay, S.A. (2013). Global approaches for telling time: Omics and the *Arabidopsis* circadian clock. *Seminars in Cell and Developmental Biology* 24: 383–392

Chow, B.Y., Helfer, A., Nusinow, D.A., and Kay, S.A. (2012). ELF3 recruitment to the PRR9 promoter requires other Evening Complex members in the *Arabidopsis* circadian clock. *Plant Signaling & Behavior* 7: 170–173

Chu, Y., and Corey, D.R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics* 22: 271–274

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., *et al*. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25 (11): 1422–1423 doi:10.1093/bioinformatics/btp163

Corellou, F., Schwartz, C., Motta, J.P., Djouani-Tahri, E.B., Sanchez, F., and Bouget, F.Y. (2009). Clocks in the Green Lineage: Comparative Functional Analysis of the Circadian Architecture of the Picoeukaryote *Ostreococcus*. *The Plant Cell Online* 21: 3436–3449

Covington, M.F., Maloof, J.N., Straume, M., Kay, S.A., and Harmer, S.L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biology* 9: pp130

Cox, R.E., Burlingame, A.L., and Wilson, D.M. (1973). Botryococcene-a Tetramethylated Acyclic Triterpenoid. *J. Chem. Soc., Chem. Commun*. 1–2

Dai, S., Wei, X., Pei, L., Thompson, R.L., Liu, Y., Heard, J.E., Ruff, T.G., and Beachy, R.N. (2011). BROTHER OF LUX ARRHYTHMO Is a Component of the *Arabidopsis* Circadian Clock. *The Plant Cell Online* 23: 961–972

Del Campo, J.A., Rodr´ıguez, H., Moreno, J., Vargas, A., Rivas, J., and Guerrero, M.G. (2001). Lutein production by *Muriellopsis sp*. in an outdoor tubular photobioreactor. *Journal of Biotechnology* 85 (3): 289–295

Deluc, L.G., Grimplet, J., Wheatley, M.D., Tillett, R.L., Quilici, D.R., Osborne, C., Schooley, D.A., Schlauch, K.A., Cushman, J.C., and Cramer, G.R. (2007) Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development. *BMC Genomics* 8: pp429

Delucchi, M.A. (2003). A lifecycle emissions model (LEM): Lifecycle emissions from transportation fuels, motor vehicles, transportation modes, electricity, use, heating and cooking fuels and materials (Institute of Transportation Studies)

Disch, A., Schwender, J., Muller, C., Lichtenthaler, H.K., and Rohmers, M. (1998). Distribution of the mevalonate and glyceraldehyde phosphate/ pyruvate pathways for isoprenoid biosynthesis in unicellular algae and the cyanobacterium *Synechocystis* PCC 6714. *Journal of Biochemistry* 333: 381–388

Ditty, J.L., Williams, S.B., and Golden, S.S. (2003). A cyanobacterial circadian timing mechanism. *Annu. Rev. Genet.* 37: 513–543

Dixon, L.E., Knox, K., Kozma-Bognár, L., Southern, M.M., Pokhilko, A., and Millar, A.J. (2011). Temporal Repression of Core Circadian Genes Is Mediated through EARLY FLOWERING 3 in *Arabidopsis*. *Current Biology* 21: 120–125

Doherty, C.J., and Kay, S.A. (2010). Circadian Control of Global Gene Expression Patterns. *Annu. Rev. Genet.* 44: 419–444

Donohue, T., and Cogdell, R. (2006). Microorganisms and clean energy. *Nature Reviews Microbiology* 4 (11): pp800

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7 (10): e1002195. doi:10.1371/journal.pcbi.1002195.g006

Eroglu, E., and Melis, A. (2010). Extracellular terpenoid hydrocarbon extraction and quantitation from the green microalgae *Botryococcus braunii* var. Showa. *Bioresource Technology* 101: 2359–2366

Eroglu, E., Okada, S., and Melis, A. (2010). Hydrocarbon productivities in different *Botryococcus* strains: comparative methods in product quantification. *Journal of Applied Phycology* 23: 763–775

Farinas, B., and Mas, P. (2011). Functional implication of the MYB transcription factor RVE8/LCL5 in the circadian control of histone acetylation. *The Plant Journal* 66: 318–329

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Research* 20: 45–58

Fowler, S., Lee, K., Onouchi, H., Samach, A., Richardson, K., Morris, B., Coupland, G., and Putterill, J. (1999). GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in *Arabidopsis* and encodes a protein with several possible membrane-spanning domains. *The EMBO Journal* 18: 4679–4688

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., *et al*. (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 10: pp161

Fuhrmann, A., Mingram, J., Lücke, A., Lu, H., Horsfield, B., Liu, J., Negendank, J.F.W., Schleser, G.H., and Wilkes, H. (2003). Variations in organic matter composition in sediments from Lake Huguang Maar (Huguangyan), south China during the last 68 ka: implications for environmental and climatic change. *Organic Geochemistry* 34: 1497–1515

Fujiwara, S., Wang, L., Han, L., Suh, S.S., Salome, P.A., McClung, C.R., and Somers, D.E. (2008). Post-translational Regulation of the *Arabidopsis* Circadian Clock through Selective Proteolysis and Phosphorylation of Pseudo-response Regulator Proteins. *Journal of Biological Chemistry* 283: 23073–23083

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., *et al*. (2009). *Nature Biotechnology* 27 (11): 1013–1023. doi:10.1038/nbt.1585

Gallagher, E., Berry, A., and Archer, G. (2008). The Gallagher review of the indirect effects of biofuels production (Renewable Fuels Agency)

Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature* 8: 469–477

Gasteiger, E. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31 (13): 3784–3788. doi:10.1093/nar/gkg563

Gendron, J.M., Pruneda-Paz, J.L., Doherty, C.J., Gross, A.M., Kang, E.S., and Kay, S.A. (2012). *Arabidopsis* circadian clock protein, TOC1, is a DNA-binding transcription factor. *PNAS* 109: 3167–3172

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32: 258–261. doi:10.1093/nar/gkh036

Giebultowicz, J.M., Stanewsky, R., Hall, J.C., and Hege, D.M. (2000). Transplanted Drosophila excretory tubules maintain circadian clock cycling out of phase with the host. *Current Biology* 10: 107–110

Glikson, M., Lindsay, K., and Saxby, J. (1989). *Botryococcus*- A planktonic green alga, the source of petroleum through the ages: Transmission electron microscopical studies of oil shales and petroleum source rocks. *Organic Geochemistry* 14: 595–608

Gouveia, L., and Oliveira, A.C. (2008). Microalgae as a raw material for biofuels production. *J Ind Microbiol Biotechnol* 36: 269–274

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., *et al*. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29 (7): 644–652. doi:10.1038/nbt.1883

Greenwell, H.C., Laurens, L.M.L., Shields, R.J., Lovitt, R.W., and Flynn, K.J. (2010). Placing microalgae on the biofuels priority list: a review of the technological challenges. *Journal of the Royal Society Interface* 7: 703–726

Grice, K., Schouten, S., Nissenbaum, A., Charrach, J., and Sinninghe Damste, J.S. (1998). A remarkable paradox: Sulfurised freshwater algal (*Botryococcus braunii*) lipids in an ancient hypersaline euxinic ecosystem. *Organic Geochemistry* 28: 195–216

Guarnieri, M. T., Nag, A., Smolinski, S. L., Darzins, A., Seibert, M., & Pienkos, P. T. (2011). Examination of Triacylglycerol Biosynthetic Pathways via *De novo* Transcriptomic and Proteomic Analyses in an Unsequenced Microalga. *PLoS ONE*, 6 (10): e25851. doi:10.1371/journal.pone.0025851.g006

Gutman, B. L. (2004). Chlamydomonas and *Arabidopsis*. A Dynamic Duo. *Plant Physiology*, 135 (2): 607–610. doi:10.1104/pp.104.041491

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38 (12): 131–131. doi:10.1093/nar/gkq224

Harmer, S.L. (2000). Orchestrated Transcription of Key Pathways in *Arabidopsis* by the Circadian Clock. *Science* 290: 2110–2113

Harmer, S.L., Panda, S., and Kay, S.A. (2001). Molecular Bases of Circadian Rhythms. *Annual Review of Cell & Developmental Biology* 17: 215–253

Haruta, M., Sabat, G., Stecker, K., Minkoff, B. B., & Sussman, M. R. (2014). A Peptide Hormone and Its Receptor Protein Kinase Regulate Plant Cell Expansion. *Science*, 343 (6169): 408–411. doi:10.1126/science.1244454

Hedges, S., Blair, J. E., Venturi, M. L., & Shoe, J. L. (2004). A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology*, 4 (1): 2. doi:10.1186/1471-2148-4-2

Helfer, A., Nusinow, D.A., Chow, B.Y., Gehrke, A.R., Bulyk, M.L., and Kay, S.A. (2011). *LUX ARRHYTHMO* Encodes a Night time Repressor of Circadian Gene Expression in the *Arabidopsis* Core Clock. *Current Biology* 21: 126–133

Henriques, R., and Mas, P. (2013). Chromatin remodeling and alternative splicing: Pre- and post-transcriptional regulation of the *Arabidopsis* circadian clock. *Seminars in Cell and Developmental Biology* 24: 399–406

Hipkins, M. F., & Baker, N. R. (1986). Photosynthesis energy transduction. A practical approach. Oxford: IRL Press.

Hsu, P.Y., Devisetty, U.K., and Harmer, S.L. (2013). Accurate timekeeping is controlled by a cycling activator in *Arabidopsis*. eLife 2.

Hu, Q.-N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X., & Deng, Z. (2012). Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. *PLoS ONE*, 7 (12): e52901. doi:10.1371/journal.pone.0052901.t003

Huang, W., Perez-Garcia, P., Pokhilko, A., Antoshechkin, I., Riechmann, J.L., and Mas, P. (2012). Mapping the core of the *Arabidopsis* Circadian Clock Defines the Network Structure of the Oscillator. *Science* 336: pp75

Huang, Y., Street-Perrott, F.A., Perrott, R.A., Metzger, P., and Eglinton, G. (1999). Glacial-interglacial environmental changes inferred from molecular and compound-specific carbon 13 analyses of sediments from Sacred Lake, Mt. Kenya. *Geochimica at Cosmochimica Acta* 63: 1383–1404

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21 (9): 1552–1560. doi:10.1101/gr.120618.111

Iancu, O.D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing RNA-Seq data for *de novo* coexpression network inference. *Bioinformatics* 28: 1592–1597

Ioki, M., Baba, M., Nakajima, N., Shiraiwa, Y., & Watanabe, M. M. (2012a). Transcriptome analysis of an oil-rich race B strain of *Botryococcus braunii* (BOT-22) by *de novo* assembly of pyrosequencing cDNA reads. Bioresource Technology, 109, 292–296. doi:10.1016/j.biortech.2011.08.104

Ioki, M., Baba, M., Nakajima, N., Shiraiwa, Y., & Watanabe, M. M. (2012b). Transcriptome analysis of an oil-rich race B strain of *Botryococcus braunii* (BOT-70) by *de novo* assembly of 5-end sequences of full-length cDNA clones. *Bioresource Technology*, 109: 277–281. doi:10.1016/j.biortech.2011.11.047

Ishimatsu, A., Matsuura, H., Sano, T., Kaya, K., and Watanabe, M.M. (2012). Biosynthesis of Isoprene Units in the C34 Botryococcene Molecule Produced by *Botryococcus Braunii* Strain Bot-22. *Procedia Environmental Sciences* 15: 56–65

Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C.R., Tanabe, A., Golden, S.S., Johnson, C.H., and Kondo, T. (1998). Expression of a Gene Cluster *kaiABC* as a Circadian Feedback Process in Cyanobacteria. *Science* 281: 1519–1523

Johnson, C.H. (2001). Endogenous timekeepers in photosynthetic organisms. *Annual Review of Physiology* 63: 695–728

Johnson, M. T. J., Carpenter, E. J., Tian, Z., Bruskiewich, R., Burris, J. N., Carrigan, C. T., *et al*. (2012). Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes. *PLoS ONE*, 7 (11): e50226. doi:10.1371/journal.pone.0050226.s011

Jonak, C. (2002). Complexity, Cross Talk and Integration of Plant MAP Kinase Signalling. *Current Opinion in Plant Biology*, 5 (5): 415–424. doi:10.1016/S1369-5266(02)00285-6

Kalra, S., Puniya, B. L., Kulshreshtha, D., Kumar, S., Kaur, J., Ramachandran, S., & Singh, K. (2008). A Perspective on the Biotechnological Potential of Microalgae. *Critical Reviews in Microbiology*, 34 (2): 77–88. doi:10.1080/10408410802086783

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42: 199–205. doi:10.1093/nar/gkt1076

Kawachi, M., Tanoi, T., Demura, M., Kaya, K., and Watanabe, M.M. (2012). Relationship between hydrocarbons and molecular phylogeny of *Botryococcus braunii*. *Algal Research* (2): 114–119

Khatri, W., Hendrix, R., Niehaus, T., Chappell, J., & Curtis, W. R. (2014). Hydrocarbon Production in High Density *Botryococcus braunii* Race B Continuous Culture. *Biotechnology and Bioengineering*, 11 (3), 1–11. doi:10.1002/bit.25126/abstract

Kikis, E.A., Khanna, R., and Quail, P.H. (2005). ELF4 is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components CCA1 and LHY. *The Plant Journal* 44: 300–313

Kircher, M., Heyn, P., & Kelso, J. (2011). Addressing challenges in the production and analysis of Illumina sequencing data. *Annual Review of Nutrition*, 12 (382): 171–201. doi:10.1186/1471-2164-12-382

Kitayama, Y., Nishiwaki, T., Terauchi, K., and Kondo, T. (2008). Dual KaiC-based oscillations constitute the circadian system of cyanobacteria. *Genes & Development* 22: 1513–1521

Kiyota, M., Numayama, N., and Goto, K. (2006). Circadian rhythms of the l-ascorbic acid level in Euglena and spinach. J*ournal of Photochemistry and Photobiology B: Biology* 84: 197–203

Koike, N., Yoo, S.H., Huang, H.C., Kumar, V., Lee, C., Kim, T.K., and Takahashi, J.S. (2012). Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. *Science* 338: 349–354

Kolmos, E., Nowak, M., Werner, M., Fischer, K., Schwarz, G., Mathews, S., Schoof, H., Nagy, F., Bujnicki, J.M., and Davis, S.J. (2009). Integrating ELF4 into the circadian system through combined structural and functional studies. *HFSP Journal* 3: 350–366

Kondo, T. (2007). A Cyanobacterial Circadian Clock Based on the Kai Oscillator. *Cold Spring Harbor Symposia on Quantitative Biology* 72: 47–55

Krinsky, N.I., Landrum, J.T., and Bone, R.A. (2003). Biologic mechanisms of the protective role of lutein and zeaxanthin in the eye. *Annu. Rev. Nutr*. 23: 171–201

Lakin-Thomas, P.L., and Brody, S. (2004). Circadian rhythms in microorganisms: New Complexities. *Annu. Rev. Microbiol.* 58: 489–519

Largeau, C., Casadevall, E., Berkaloff, C., and Dhamelincourt, P. (1980). Sites of accumulation and composition of hydrocarbons in *Botryococcus braunii*. *Phytochemisty* 19: 1043–1051

Lee, B.D., Kim, S.B., Kwon, G.S., Yoon, B.D., Yoon, B.D. (1998). Effects of harvesting method and growth stage on the flocculation of the green alga *Botryococcus braunii*. *Letters in Applied Microbiology*, 27: 14-18

Lehto, J., Oasmaa, A., Solantausta, Y., Kytö, M., and Chiaramonti, D. (2014). Review of fuel oil quality and combustion of fast pyrolysis bio-oils from lignocellulosic biomass. *Applied Energy* 116: 178–190

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12 (1): 323. doi:10.1186/1471-2105-12-323

Lichtenthaler, H.K. (1999). The 1-deoxy-d-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annu Rev Plant Physiol Plant Mol Biol* 50: 47–65

Lichtenthaler, H.K. (2014). Treatment of industrial wastewaters by microalgal bacterial flocs in sequencing batch reactors. *Bioresource Technology* 161: 245–254

Liu, Y., Tsinoremas, N.F., Johnson, C.H., Lebedeva, N.V., Golden, S.S., Ishiura, M., and Kondo, T. (1995). Circadian orchestration of gene expression in cyanobacteria. *Genes & Development* 9: 1469–1478

Locke, J.C.W., Kozma-Bognár, L., Gould, P.D., Fehér, B., Kevei, É., Nagy, F., Turner, M.S., Hall, A.T., and Millar, A.J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Mol Syst Biol* 2: pp59

Locke, J.C.W., Southern, M.M., Kozma-Bognár, L., Hibberd, V., Brown, P.E., Turner, M.S., and Millar, A.J. (2005). Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol Syst Biol* 1

Lohr, S. (2011). Jobs Tried Exotic Treatments to Combat Cancer, Book Says. New York Times

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., & Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40 (1): 622–627. doi:10.1093/nar/gks540

Loivamaki, M., Louis, S., Cinege, G., Zimmer, I., Fischbach, R.J., and Schnitzler, J.P. (2006). Circadian Rhythms of Isoprene Biosynthesis in Grey Poplar Leaves. *Plant Physiology* 143: 540–551

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15 (12): 31. doi:10.1186/s13059-014-0550-8

Lowrey, P.L., and Takahashi, J.S. (2004). Mammalian circadian biology: Elucidating Genome-Wide Levels of Temporal Organization. *Annu. Rev. Genom. Human Genet.* 5: 407–441

Lupi, F. M., Fernandes, H. M. L., Sa-Correia, I., & Novais, J. M. (1991). Temperature profiles of cellular growth and exopolysaccharide synthesis by *Botryococcusbraunii* Kutz. UC 58. *Joural of Applied Phycology*: 3, 35–42

Lyubetsky, V. A., Seliverstov, A. V., & Zverkov, O. A. (2013). Transcription Regulation of Plastid Genes Involved in Sulfate Transport in Viridiplantae. *BioMed Research International*, 2013 (1): 1–6. doi:10.1098/rstb.2002.1187

Manning, G. (2002). The Protein Kinase Complement of the Human Genome. *Science*,

298 (5600), 1912–1934. doi:10.1126/science.1075762

Mardis, E.R. (2008). Next-Generation DNA Sequencing Methods. *Annu. Rev. Genom. Human Genet.* 9: 387–402

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18: 1509–1517

Martin, J. A., & Wang, Z. (2011).Next-generation transcriptome assembly. *Nature*, 12 (10): 671–682. doi:10.1038/nrg3068

Mata, T.M., Martins, A.A., and Caetano, N.S. (2010). Microalgae for biodiesel production and other applications: A review. *Renewable and Sustainable Energy Reviews* 14: 217–232

Matsukawa, R., Hotta, M., Masuda, Y., Chihara, M., and Karube, I. (2000). Antioxidants from carbon dioxide fixing *Chlorella sorokiniana*. *Journal of Applied Phycology* 12: 263–267

Matsushika, A., Makino, S., Kojima, M., and Mizuno, T. (2000). Circadian Waves of Expression of the APRR1/TOC1 Family of Pseudo-Response Regulators in *Arabidopsis thaliana*: Insight into the Plant Circadian Clock. *Plant Cell Physiology* 41: 1002–1012

Matsushima, D., Jenke-Kodama, H., Sato, Y., Fukunaga, Y., Sumimoto, K., Kuzuyama, T., *et al*. (2012). The single cellular green microalga *Botryococcus braunii*, race B possesses three distinct 1-deoxy-d-xylulose 5-phosphate synthases. *Plant Science*, 185-186, 309–320. doi:10.1016/j.plantsci.2012.01.002

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. *PNAS* 74: 560–564

Merino, E., Balbas, P., Puente, J. L., & Bolivar, F. (1994). Antisense Overlapping Open Reading Frames in Genes From Bacteria to Humans. *Nucleic Acids Research*, 22 (10): 1903–1908

Metzger, P., Allard, B., and Casadevall, E. (1990). Structure and chemistry of a new chemical race of *Botryococcus braunii* (Chlorophycae) that produces lycopadiene, a tetraterpenoid hydrocarbon. *Journal of Phycology* 25: 258–266

Metzger, P., and Casadevall, E. (1987). Lycopadiene, a tetraterpenoid hydrocarbon from new strains of the green alga *Botryococcus braunii*. *Tetrahedron Letter* 28: 3931–3934

Metzger, P., Berkaloff, C., Casadevall, E., & Coute, A. (1985). Alkadiene- and botryococcene-producing races of wild strains of *Botryococcus braunii*. *Phytochemisty*, 24 (10): 2305–2312

Metzger, P., Casadevall, E., and Coute, A. (1988). Botryococcene distribution in strains of the green alga *Botryococcus braunii*. *Phytochemistry* 27: 1388–1988

Metzger, P., Casadevall, E., Pouet, M.J., and Pouet, Y. (1985a). Structures of some botryococcenes: Branched hydrocarbons from the B-race of the green alga *Botryococcus braunii*. *Phytochemisty* 24: 2995–3002

Metzger, P., David, M., and Casadevall, E. (1987). Biosynthesis of triterpenoid hydrocarbons in the B-race of the green alga *Botryococcus braunii*. Sites of production and nature of the methylating agent. *Phytochemistry* 26: 129–134

Metzker, M. L. (2009). Sequencing technologies- the next generation. Nature, 11 (1): 31–46. doi:10.1038/nrg2626

Mirón, A.S., García, M.C.C., Gómez, A.C., Camacho, F.G., Grima, E.M., and Chisti, Y. (2003). Shear stress tolerance and biochemical characterization of *Phaeodactylum tricornutum* in quasi steady-state continuous culture in outdoor photobioreactors. *Biochemical Engineering Journal* 16, 287–297

Mittag, M. (2001). Circadian rhythms in microalgae. *International Review Cytology* 206: 213–247

Mittag, M. (2005). The Circadian Clock in *Chlamydomonas reinhardtii*. What Is It For? What Is It Similar To? *Plant Physiology*, 137 (2): 399–409. doi:10.1104/pp.104.052415

Mittag, M., and Wagner, V. (2005). The circadian clock of the unicellular eukaryotic model organism *Chlamydomonas reinhardtii*. *Biological Chemistry* 384

Moldowan, J.M., and Seifert, W.K. (1980). First discovery of botryococcane in petroleum. *J. Chem. Soc., Chem. Commun.* 912

Molnár, I., Lopez, D., Wisecaver, J. H., Devarenne, T. P., Weiss, T. L., Pellegrini, M., & Hackett, J. D. (2012). Bio-crude transcriptomics: Gene discovery and metabolic network reconstruction for the biosynthesis of the terpenome of the hydrocarbon oil-producing green alga, *Botryococcusbraunii* race B (Showa)*. BMC Genomics, 13(1), 576. doi:10.1186/1471-2164-13-576

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* 24: 22–30

Nagel, D.H., and Kay, S.A. (2012). Complexity in the Wiring and Regulation of Plant Circadian Networks. *Current Biology* 22: 648–657

Nakamichi, N. (2005). PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, Together Play Essential Roles Close to the Circadian Clock of *Arabidopsis thaliana*. *Plant and Cell Physiology* 46: 686–698

Nakamichi, N. (2011). Molecular Mechanisms Underlying the *Arabidopsis* Circadian Clock. Plant and Cell Physiology 52, 1709–1718

Neori, A. (2010). “Green water” microalgae: the leading sector in world aquaculture. *Journal of Applied Phycology* 23: 143–149

Niehaus, T. D., Okada, S., Devarenne, T. P., Watt, D. S., Sviripa, V., & Chappell, J. (2011). Identification of unique mechanisms for triterpene biosynthesis in *Botryococcus braunii*. *PNAS*, 108 (30): 1–12. doi:10.1073/pnas.1106222108/-/DCSupplemental

Niehaus, T.D., Kinison, S., Okada, S., Yeo, Y.S., Bell, S.A., Cui, P., Devarenne, T.P., and Chappell, J. (2012). Functional Identification of Triterpene Methyltransferases from *Botryococcus braunii* Race B. *Journal of Biological Chemistry* 287: 8163–8173

Nishida, E., & Gotoh, Y. (1993). THe MAP kinase cascade is essential for diverse signal transduction pathways. *Trends in Biochemical Science*, 18: 128–131

Okada, S., Devarenne, T. P., & Chappell, J. (2000). Molecular Characterization of Squalene Synthase from the Green Microalga *Botryococcus braunii*, Race B. *Archives of Biochemistry and Biophysics*, 373 (2): 307–317. doi:10.1006/abbi.1999.1568

Ozsolak, F., & Milos, P. M. (2010). RNA sequencing: advances, challenges and opportunities. *Nature* 12 (2): 87–98. doi:10.1038/nrg2934

Pajuelo, E., Pajuelo, P., Clemente, M.T., and Marquez, A.J. (1995). Regulation of the expression of ferredoxin-nitrite reductase in synchronous cultures of *Chlamydomonas reinhardtii*. *Biochimica Et Biophysica Acta* 1249: 72–78

Pittendrigh, C.S. (1993). Temporal organization: Reflections of a Darwinian Clock-Watcher. *Annual Review of Physiology* 55: 17–54

Plautz, J.D. (1997). Independent Photoreceptive Circadian Clocks Throughout Drosophila. *Science* 278: 1632–1635

Pokhilko, A., Fernandez, A.P., Edwards, K.D., Southern, M.M., Halliday, K.J., and Millar, A.J. (2012). The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. *Mol Syst Biol* 8 (574)

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., *et al*. (2013). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, 42: 756–763. doi:10.1093/nar/gkt1114

Pruneda-Paz, J.L., Breton, G., Para, A., and Kay, S.A. (2009). A Functional Genomics Approach Reveals CHE as a Component of the *Arabidopsis* Circadian Clock. *Science* 323: 1481–1485

Raja, R., Hemaiswarya, S., Kumar, N.A., Sridhar, S., and Rengasamy, R. (2008). A Perspective on the Biotechnological Potential of Microalgae. *Critical Reviews in Microbiology* 34: 77–88

Rawat, R., Schwartz, J., Jones, M.A., Sairanen, I., Cheng, Y., Andersson, C.R., Zhao, Y., Ljung, K., and Harmer, S.L. (2009). REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways. *PNAS* 106: 16883–16888

Rismani-Yazdi, H., Haznedaroglu, B. Z., Bibby, K., & Peccia, J. (2011). Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next- generation biofuels. *BMC Genomics*, 12 (1): 148. doi:10.1186/1471-2164-12-148

Rivas, M. O., Vargas, P., & Riquelme, C. E. (2010). Interactions of *Botryococcus braunii* Cultures with Bacterial Biofilms. *Microbial Ecology*, 60 (3): 628–635. doi:10.1007/s00248-010-9686-6

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.

Bioinformatics, 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Rodolfi, L., Chini Zittelli, G., Bassi, N., Padovani, G., Biondi, N., Bonini, G., and Tredici, M.R. (2009). Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. *Biotechnology and Bioengineering* 102: 100–112

Rosato, E., and Kyriacou, C.P. (2002). Origins of Circadian Rhythmicity. *Journal of Biological Rhythms* 17: 506–511

Rugnone, M.L., Soverna, A.F., Sanchez, S.E., Schlaen, R.G., Hernando, C.E., Seymour, D.K., Mancini, E., Chernomoretz, A., Weigel, D., Mas, P., *et al*. (2013). LNK genes integrate light and clock signaling networks at the core of the *Arabidopsis* oscillator. *PNAS* 110: 12120–12125

Salome, P.A. (2005). *PSEUDO-RESPONSE REGULATOR 7* and *9* Are Partially Redundant Genes Essential for the Temperature Responsiveness of the *Arabidopsis* Circadian Clock. *The Plant Cell Online* 17, 791–803

Salvador, M.L., Klein, U., and Bogorad, L. (1998). Endogenous Fluctuation of DNA topology in the Chloroplast of *Chlamydomonas reinhardtii*. *Molecular and Cellular Biology* 18: 7235–7242

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS* 74: 5463–5467

Sasso, S., Pohnert, G., Lohr, M., Mittag, M., and Hertweck, C. (2011). Microalgae in the postgenomic era: a blooming reservoir for new natural products. *FEMS Microbiol Rev* 36: 761–785

Sato, Y., Ito, Y., Okada, S., Murakami, M., and Abe, H. (2003). Biosynthesis of the triterpenoids, botryococcenes and tetramethylsqualene in the B race of *Botryococcus braunii* via the non-mevalonate pathway. *Tetrahedron Letters* 44: 7035–7037

Sawayama, S., Inoue, S., and Yokoyama, S.-Y. (1995). Phylogenetic position of *Botryococcus braunii* (Chlorophyceae) based on small subunit ribosomal RNA sequence data. *Journal of Phycology* 31: 419–420

Scarlet, N., Dallemand, J.F., and Gallego Pinilla, F. (2008). Impact on agricultural land resources of biofuels production and use in the European Union. pp. 1–10

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M., and Wisman, E. (2001). Microarray analysis of diurnal and circadian- regulated genes in *Arabidopsis*. *The Plant Cell* 13: 113–123

Schmid, R., & Blaxter, M. L. (2008). annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, 9 (1), 180. doi:10.1186/1471-2105-9-180

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., *et al*. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, 7 (1), 3. doi:10.1186/1471-2199-7-3

Schwender, J., Gemunden, C., and Lichtenthaler, H.K. (2001). Chlorophyta exclusively use the 1-deoxyxylulose-5-phosphate. *Planta* 212, 416–423

Schwender, J., Seemann, M., Lichtenthaler, H.K., and Rohmers, M. (1996). Biosynthesis of isoprenoids (carotenoids, sterols, prenyl side-chains of chlorophylls and plastoquinone) via a novel pyruvate/ glyceraldehyde-3-phosphate non-mevalonate pathway in the green alga *Scenedesmus obliquus*. *Journal of Biochemistry* 316: 73–80

Senousy, H.H., Beakes, G.W., and Hack, E. (2004). Phylogenetic Placement of *Botryococcus braunii* (Trebouxiophyceae) and *Botryococcus* Sudeticus Isolate Utex 2629 (Chlorophyceae). *Journal of Phycology* 40: 412–423

Sheehan, J., Dunahay, T., Benemann, J., and Roessler, P. (2003). A Look Back at the U.S Department of Energy's Aquatic Species Program: Biodiesel from Algae (Golden).

Sheih, I.C., Fang, T.J., and Wu, T.-K. (2009). Isolation and characterisation of a novel angiotensin I-converting enzyme (ACE) inhibitory peptide from the algae protein waste. *Food Chemistry* 115: 279–284

Simkin, A.J. (2004). Circadian Regulation of the PhCCD1 Carotenoid Cleavage Dioxygenase Controls Emission of -Ionone, a Fragrance Volatile of Petunia Flowers. *Plant Physiology* 136: 3504–3514

Spolaore, P., Joannis-Cassan, C., Duran, E., and Isambert, A. (2006). Commercial applications of microalgae. *Journal of Bioscience and Bioengineering* 101: 87–96

Stephens, E., Ross, I.L., King, Z., Mussgnug, J.H., Kruse, O., Posten, C., Borowitzka, M.A., and Ben Hankamer (2010). An economic and technical evaluation of microalgal biofuels. *Nature Biotechnology* 28: 126–128

Stirnimann, C. U., Petsalaki, E., Russell, R. B., & Müller, C. W. (2010). WD40 proteins propel cellular networks. *Trends in Biochemical Sciences*, 35 (10): 565–574 doi:10.1016/j.tibs.2010.04.003

Strayer, C., Oyama, T., Schultz, T.F., Raman, R., Somers, D.E., Mas, P., Panda, S., Kreps, J.A., and Kay, S.A. (2000). Cloning of the *Arabidopsis* Clock Gene *TOC1*, an Autoregulatory Respomse Regulator Homolog. Science 289, 768–770.

Strickfaden, S. C., Winters, M. J., Ben-Ari, G., Lamson, R. E., Tyers, M., & Pryciak, P. M. (2007). A Mechanism for Cell-Cycle Regulation of MAP Kinase Signaling in a Yeast Differentiation Pathway. *Cell*, 128 (3): 519–531. doi:10.1016/j.cell.2006.12.032

Sun, D., Zhu, J., Fang, L., Zhang, X., Chow, Y., & Liu, J. (2013). *De novo* transcriptome profiling uncovers a drastic downregulation of photosynthesis upon nitrogen deprivation in the nonmodel green alga *Botryosphaerella sudeticus*. *BMC Genomics*, 14 (1): 1–1. doi:10.1186/1471-2164-14-715

Tanabe, Y., Ioki, M., & Watanabe, M. M. (2013). The fast-growing strain of hydrocarbon-rich green alga *Botryococcus braunii*, BOT-22, is a vitamin B12 autotroph. *Journal of Applied Phycology*, 26 (1): 9–13. doi:10.1007/s10811-013-0045-0

Tanabe, Y., Kato, S., Matsuura, H., & Watanabe, M. M. (2012). A *Botryococcus* strain with bacterial ectosymbionts grows fast and produces high amount of hydrocarbons. *Procedia Environmental Sciences*, 45: 22–26. doi:10.1016/j.proenv.2012.05.005

Tang, Q., Ma, X., Mo, C., Wilson, I.W., Song, C., Zhao, H., Yang, Y., Fu, W., and Qiu,

D. (2011). An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. *BMC Genomics* 12

Temperley, B.N. (1936). *Botryococcus* and the algal coals. Part II. The boghead controversy and the morphology of the boghead algae. *Transactions of the Royal Society of Edinburgh* 58: 855–868

Tomita, J., Nakajima, M., Kondo, T., and Iwasaki, H. (2005). No Transcription-Translation Feedback in Circadian Rhythm of KaiC Phosphorylation. *Science* 307: 251–254

Tyson, R. (1995a). Sedimentary Organic Matter. In Organic Facies and Palynofacies, (London: Chapman & Hall), p. 640

Ugwu, C.U., Aoyagi, H., and Uchiyama, H. (2008). Photobioreactors for mass cultivation of algae. *Bioresource Technology* 99: 4021–4028

Vazquez-Duhalt, R., and Greppin, H. (1987). Growth and production of cell constituents in batch cultures of *Botryococcus* sudeticus. *Phytochemistry* 26: 885–890

Volkman, J.K. (2014). Acyclic isoprenoid biomarkers and evolution of biosynthetic pathways in green microalgae of the genus *Botryococcus*. *Organic Geochemistry* 75: 36–47

Wake, L.V., and Hillen, L.W. (1980). Study of a “bloom” of the oil-rich alga *Botryococcus braunii* in the Darwin River Reservoir. *Biotechnology and Bioengineering* 22: 1637–1656

Wake, L.V., and Hillen, L.W. (1981). Nature and hydrocarbon content of blooms of the alga *Botryococcus braunii* occurring in Australian freshwater lakes. *Australian Journal of Marine and Freshwater Research* 32: 353–367

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476

Wang, Z.-Y., and Tobin, E.M. (1998). Constitutive Expression of the *CIRCADIAN ASSOCIATED 1* (*CCA1*) Gene Disrupts Circadian Rhythms and Suppresses Its Own Expression. *Cell* 26: 1207–1217

Weiss, T.L., Chun, H.J., Okada, S., Vitha, S., Holzenburg, A., Laane, J., and Devarenne, T.P. (2010). Raman Spectroscopy Analysis of Botryococcene Hydrocarbons from the Green Microalga *Botryococcus braunii*. *Journal of Biological Chemistry* 285: 32458–32466

Weiss, T.L., Roth, R., Goodson, C., Vitha, S., Black, I., Azadi, P., Rusch, J., Holzenburg, A., Devarenne, T.P., and Goodenough, U. (2012). Colony Organization in the Green Alga *Botryococcus braunii* (Race B) Is Specified by a Complex Extracellular Matrix. *Eukaryotic Cell* 11: 1424–1440

Welsh, D.K., Logothetis, D.E., Meister, M., and Reppert, S.M. (1995). Individual Neurons Dissociated from Rat Suprachiasmatic Nucleus Express Independently Phased Circadian Firing Rhythms. *Neuron* 14: 697–706

Wijnen, H., and Young, M.W. (2006). Interplay of Circadian Clocks and Metabolic

Rhythms. *Annu. Rev. Genet.* 40: 409–448

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239–1243

Wolf, F.R., Nonomura, A.M., and Bassham, J. (1985). Growth and branched hydrocarbon production in a strain of *Botryococcus braunii* (Chlorophyta). *Journal of Phycology* 21: 388–396

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21 (9): 1859–1875. doi:10.1093/bioinformatics/bti310

Xu, C., & Min, J. (2011). Structure and function of WD40 domain proteins. *Protein & Cell*, 2 (3): 202–214. doi:10.1007/s13238-011-1018-1

Xu, Y., Mori, T., and Johnson, C.H. (2000). Circadian clock-protein expression in cyanobacteria: rhythms and phase setting. *The EMBO Journal* 19: 3349–3357

Yang, S., Guarnieri, M. T., Smolinski, S., Ghirardi, M., & Pienkos, P. T. (2013). *De novo* transcriptomic analysis of hydrogen production in the green alga *Chlamydomonas moewusii* through RNA-Seq. *Biotechnology for Biofuels*, 6 (1): 1–1. doi:10.1186/1754-6834-6-118

Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G.P., Luo, S., Khrebtukova, I., Gnirke, A., *et al*. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *PNAS* 3264–3269

Yoon, H. S. (2004). A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Molecular Biology and Evolution*, 21 (5): 809–818. doi:10.1093/molbev/msh075

Young, M.W., and Kay, S.A. (2001). Time Zones: A comparative Genetics of Circadian Clocks. *Nature Reviews Genetics* 2: 702–715

Zeilinger, M.N., Farré, E.M., Taylor, S.R., Kay, S.A., and Doyle, F.J. (2006). A novel computational model of the circadian clock in *Arabidopsis* that incorporates PRR7 and PRR9. *Mol Syst Biol* 2

Zhang, Z., Metzger, P., and Sachs, J.P. (2007). Biomarker evidence for the co-occurrence of three races (A, B and L) of *Botryococcus braunii* in El Junco Lake, Galápagos. *Organic Geochemistry* 38: 1459–1478

Zhao, B., Schneid, C., Iliev, D., Schmidt, E.M., Wagner, V., Wollnik, F., and Mittag, M. (2004). The Circadian RNA-Binding Protein CHLAMY 1 Represents a Novel Type Heteromer of RNA Recognition Motif and Lysine Homology Domain-Containing Subunits. *Eukaryotic Cell* 3: 815–825

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Da Luo, Li, X., & Hao, P. (2011). Optimising *de novo* transcriptome assemby from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12. doi:10.1186/1471-2105-12-S14-S2

Zwicker, D., Lubensky, D.K., and Wolde, T.R.P. (2010). Robust circadian clocks from coupled protein- modification and transcription–translation cycles. 107: 22540–22545