Deposit & Copying of Dissertation Declaration

Please note that you will also need to bind a copy of this Declaration into your final, hardbound copy of thesis - this has to be the very first page of the hardbound thesis.

<table>
<thead>
<tr>
<th></th>
<th>Surname (Family Name)</th>
<th>Forenames(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kudahl</td>
<td>Ulrich Johan</td>
<td>Mr</td>
</tr>
</tbody>
</table>

A computational biology approach to studying algae-bacterial interactions

In accordance with the University Regulations in Statutes and Ordinances for the PhD, MSc and MLitt Degrees, I agree to deposit one print copy of my dissertation entitled above and one print copy of the summary with the Secretary of the Board of Graduate Studies who shall deposit the dissertation and summary in the University Library under the following terms and conditions:

1. Dissertation Author Declaration

I am the author of this dissertation and hereby give the University the right to make my dissertation available in print form as described in 2. below.

My dissertation is my original work and a product of my own research endeavours and includes nothing which is the outcome of work done in collaboration with others except as declared in the Preface and specified in the text. I hereby assert my moral right to be identified as the author of the dissertation.

The deposit and dissemination of my dissertation by the University does not constitute a breach of any other agreement, publishing or otherwise, including any confidentiality or publication restriction provisions in sponsorship or collaboration agreements governing my research or work at the University or elsewhere.

2. Access to Dissertation

I understand that one print copy of my dissertation will be deposited in the University Library for archival and preservation purposes, and that, unless upon my application restricted access to my dissertation for a specified period of time has been granted by the Board of Graduate Studies prior to this deposit, the dissertation will be made available by the University Library for consultation by readers in accordance with University Library Regulations and copies of my dissertation may be provided to readers in accordance with applicable legislation.

<table>
<thead>
<tr>
<th></th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td>12/06-17</td>
</tr>
</tbody>
</table>

Corresponding Regulation

Before being admitted to a degree, a student shall deposit with the Secretary of the Board one copy of his or her hard-bound dissertation and one copy of the summary (bearing student's name and thesis title), both the dissertation and the summary in a form approved by the Board. The Secretary shall deposit the copy of the dissertation together with the copy of the summary in the University Library where, subject to restricted access to the dissertation for a specified period of time having been granted by the Board of Graduate Studies, they shall be made available for consultation by readers in accordance with University Library Regulations and copies of the dissertation provided to readers in accordance with applicable legislation.
“A computational biology approach to studying algae-bacterial interactions”

Ulrich Johan Kudahl
Jesus College
University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

June 2018
Abstract

“A computational approach to studying algae-bacterial interactions”
By Ulrich Johan Kudahl

Microalgae have a profound effect on the world due to their large contribution to net carbon fixation. Although they are phototrophic, more than 50% of microalgae are thought to depend on external supply of metabolites such as B-vitamins. In oceans, algae are therefore often found together with a community of bacteria and form intricate networks where metabolites are exchanged. Currently, only a fraction of the related mechanisms and metabolite exchanges between algae and bacteria have been uncovered and many more are likely to exist. The work presented in this thesis is based on a model system for algae-bacterial interactions made up of the green alga, *Lobomonas rostrata* and the alphaproteobacterium *Mesorhizobium loti*. In the model system, it is known that the bacterium provides vitamin B12 to the alga and itself, whilst the alga provides fixed carbon. I have applied methods from the field of computational biology to study the interactions between these organisms and other similar partnerships, with the aim of uncovering new insights. The thesis is made up of three research chapters, each focused on using a specific method to study algae-bacterial interactions.

I developed a genome scale metabolic model of metabolism of *M. loti* that enabled simulation of growth. The model simulates 1908 enzymatic reactions and takes 1804 metabolites into account. Using the model, I simulated growth of the bacterium on 1018 different substrates with the aim of identifying substrates supplied by *L. rostrata* when the two organisms are co-cultured. In addition, I carried out a set of simulations studying the bacterium’s ability to produce B12 from 1368 different substrates. The modelling efforts in this project was
successful in enabling simulations, but it was not possible to validate the simulations with experimental data.

A transcriptomics experiment was undertaken with the aim of identifying genes related to the interaction between *L. rostrata* and *M. loti*. In the experiment, the partners from the model system was grown in axenic and co-culture conditions and RNA samples were taken from each state. Using RNA-seq, the RNA samples were sequenced and from this a candidate transcriptome was created. The expression of each putative gene was then quantified and differentially expressed genes were identified. Based on sequence similarity, candidate functions were assigned where possible. In the analysis of differentially expressed genes, it was found that there appears to be an increased expression of a transporter responsible for uptake of the plant hormone, auxin.

Currently, only a small fraction of all bacteria has been shown to produce B12 and it is not clear in which phylogenetic groups this is a common trait. I therefore applied methods from comparative genomics to study the synthesis of this metabolite in more than 8000 bacterial species. This involved developing a computational framework that allowed me to search for the presence of more than 50 genes in more than 8000 genomes in a rapid manner. I found that 37.2% of bacteria can synthesise B12 and that this capability is very common in some phylogenetic groups such as *Cyanobacteria*, but extremely rare in others such as *Lactobacillus*. I was also able to confirm that cyanobacteria are not able to make cobalamin, a variant of B12 used by eukaryotic algae, and thus they are unlikely to support algal growth in the photic zone.

In the final section of the thesis, I discuss the application of computational biology methods in this field and summarise my experience from applying genome scale modelling, comparative genomics and transcriptomics to study algae-bacterial interactions.
Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.
Acknowledgements

I would like to thank my supervisor Professor Alison Smith for accepting me in her research lab and supporting me over the past years. During my time in Cambridge, Alison’s support and advice has been invaluable and this thesis is a result of our strong collaboration over the years. Her ability to teach, mentor and motivate has been central to my development as a researcher and countless times has she proved that she is dedicated to getting the best out every student that goes through her lab.

I would also like to thank the Photo.Comm (http://photocomm.ku.dk/) training network for young researchers. Through this network, I have been able to exchange ideas, methods and experiences with other PhD students working in the same field. I have benefitted from the training courses arranged for the network in intellectual property rights, presentation skills, transcriptomics and scientific writing. The funding for this project was also related to Photo.Comm. The thesis presented here and the research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° 317184. This material reflects only the author’s views and the Union is not liable for any use that may be made of the information contained therein.

I am grateful for the collaborations I have had throughout my research and without whom this thesis would not have been the same. I would like to thank all the members of Lab 220 over the years with whom I have shared endless cups of 10’o’clock tea and discussed algal research. I would especially like to thank Katherine, Chris, Freddy, Elena, Vai and Matt for our many discussions about
mutualism and Vitamin B12 over the years and for your support. Chapter 2 in this thesis would not have been possible without the support and advice on genome scale metabolic modelling and flux balance analysis from Jun.-Prof. Dr. Oliver Ebenhöh from University of Dusseldorf and Dr. Thomas Pfau from University of Luxembourg on. Related to the study described in chapter 4 I received advice and guidance from Dr. Krys Kelly and Chris Boursnell from the Department of Plant Sciences, University of Cambridge on transcriptomics analysis and de novo assembly. I received additional support for this project during a research stay at University of Freiburg, where I was hosted by Dr. Björn Voß and worked with his group.

During my time in Cambridge, I have been part of the Jesus College community of graduate students and the Jesus College Boat Club. I am very thankful for the friendships that I made during my time at the college and for the many great outings in the boat on the river Cam.

Finally, I would like to dedicate this thesis to the love of my life, Mette. You have always provided words of encouragement when I needed it most and you have always believed in me even when I have doubted myself. This thesis would not have been possible without you and I am eternally thankful for your support.
ABSTRACT .......................................................................................................................... I

DECLARATION ................................................................................................................... III

ACKNOWLEDGEMENTS ..................................................................................................... IV

LIST OF FIGURES ............................................................................................................... X

LIST OF TABLES .................................................................................................................. XIV

ABBREVIATIONS ................................................................................................................ XVII

1 INTRODUCTION.................................................................................................................. 1

1.1 MICROBIAL COMMUNITIES.......................................................................................... 1

1.1.1 OCEAN MICROBIOME ......................................................................................... 1

1.1.2 HUMAN GUT MICROBIOME ................................................................................. 3

1.2 MICROALGAE............................................................................................................... 5

1.2.1 MICROALGAE AND THEIR GLOBAL IMPLICATIONS........................................... 5

1.2.2 ALGAE-BACTERIAL INTERACTIONS ................................................................. 6

1.3 VITAMIN B12 ............................................................................................................. 11

1.3.1 VITAMIN B12 AS AN ENZYMATIC CO-FACTOR ................................................ 11

1.3.2 VITAMIN B12 DEPENDENCE IN ALGAE .......................................................... 14

1.4 OTHER B-VITAMINS ................................................................................................... 15

1.4.1 VITAMIN B7 (BIOTIN) ....................................................................................... 15

1.4.2 VITAMIN B3 (NIACIN) ...................................................................................... 17

1.5 AIM AND STRUCTURE OF THIS THESIS ................................................................ 21

1.6 PUBLICATIONS ........................................................................................................... 22

2 GENOME SCALE METABOLIC MODEL OF M. LOTI ....................................................... 24

2.1 INTRODUCTION .......................................................................................................... 24

2.1.1 GSM MODELS AND FLUX BALANCE ANALYSIS ............................................. 27

2.1.2 ALGAE-BACTERIA MODEL SYSTEM .............................................................. 29

2.1.3 AIM OF THIS STUDY ......................................................................................... 31

2.2 METHODS ................................................................................................................... 31

2.2.1 STRAINS AND GENOME SEQUENCES ............................................................. 31

2.2.2 CREATION OF A GENOME SCALE METABOLIC RECONSTRUCTION ........... 31

2.2.3 IN SILICO EXPERIMENTS ................................................................................. 37
7.3  INSIGHTS INTO THE RED ALGAE AND EUKARYOTIC EVOLUTION FROM THE GENOME OF
PORPHYRA UMBILICALIS ................................................................. 183

7.4  QUANTITATIVE PROTEOMICS OF A B12-DEPENDENT ALGA GROWN IN CO-CULTURE
WITH BACTERIA REVEALS METABOLIC TRADE-OFFS REQUIRED FOR MUTUALISM ........... 194
List of Figures

FIGURE 1.1 – OVERVIEW OF ALGAL PHYLOGENY FROM ANCIENT ANCESTRAL ORGANISMS TO CURRENT SPECIES. FIGURE FROM COOPER & SMITH, (2015) 5

FIGURE 1.2 – EXAMPLES OF METABOLITES EXCHANGED BETWEEN ALGAE AND BACTERIA FOUND IN PREVIOUS STUDIES. NOT ALL OF THE EXAMPLES ARE FOUND TO BE EXCHANGED BETWEEN ALL ALGAE AND BACTERIA PAIRS. 10

FIGURE 1.3 - STRUCTURE OF COBALAMIN. B12 IS A CORRIN RING WITH A COBALT ION AT THE CENTRE AND NUCLEOTIDE LOOP. THE LOWER AXIAL LIGAND DMB IS SHOWN IN MARKED BOX. THE FIGURE IS ADAPTED FROM TAGA & WALKER, 2008 12

FIGURE 1.4 – OVERVIEW OF TWO ENZYMATIC REACTIONS THAT CAN BE CATALYSED BY EITHER B12-DEPENDENT OR B12-DEPENDENT ENZYMES. THE B12-INDEPENDENT FORMS ARE SHOWN IN ORANGE, WHILE THE B12-DEPENDENT FORMS ARE IN BLUE. BOTH PROCESSES ARE THOUGHT TO BE ESSENTIAL IN ALL ORGANISMS. 14

FIGURE 1.5 – A) STRUCTURE OF BIOTIN (HASTINGS ET AL, 2013) B) BIOTIN SYNTHESIS PATHWAY FROM THE PRECURSOR PIMELOYL-COA TO THE FINAL FORM OF BIOTIN. GENES FROM PLANTS/ALGAE ARE IN GREEN, WHILE GENES FROM BACTERIA IN BLUE. 17

FIGURE 1.6 – NIACIN SYNTHESIS PATHWAY. NIACIN CAN BE SYNTHESISED FROM EITHER ASPARTATE OR TRYPTOPHAN THROUGH A SERIES OF ENZYMATIC STEPS. THERE ARE TWO VARIANTS OF THE TRYPTOPHAN PATHWAY. THE ENZYME NAMES ARE MENTIONED AS WELL AS THE INTERMEDIATES IN THE SYNTHESIS PATHWAY. 20

FIGURE 2.1 – EXAMPLE OF A METABOLIC MODEL WITH THREE METABOLITES AND SEVEN REACTIONS. 28

FIGURE 2.2 – EXAMPLE OF FLUX DISTRIBUTIONS WITH CONSTRAINTS. 29

FIGURE 2.3 – EQUILIBRIUM AND B12 PRODUCTION IN THE MODEL SYSTEM. 31

FIGURE 2.4 – STOICHIOMETRIC BALANCING OF A REACTION. IN THE EXAMPLE A METABOLITE DOES NOT HAVE A CHEMICAL FORMULA. IN THIS CASE, THE
CHEMICAL FORMULA OF THE QUINONE IS REPRESENTED AS R₁ AND R₂H₂ IN THE REDUCED FORM.

FIGURE 2.5 – EXAMPLE OF A CYCLE OF REACTIONS VIOLATING THERMODYNAMICS.

FIGURE 2.6 – ANALYSIS OF B12 YIELD FROM VARIOUS SUBSTRATES

FIGURE 3.1 – SCHEMATIC OVERVIEW OF THE TWO B12 SYNTHESIS PATHWAYS.

FIGURE 3.2 – STRUCTURES OF COBALAMIN AND PSEUDOCOBALAMIN. THE LOWER LIGAND IS HIGHLIGHTED IN THE COLOURED BOX. FOR COBALAMIN THIS LOWER LIGAND IS DMB, WHILE IN PSEUDOCOBALAMIN, THE LIGAND IS REPLACED BY ADENINE.

FIGURE 3.3 – RAND ACCURACY OF PREDICTIONS BASED ON RANGE OF CUT-OFFS FOR POSITIVE CLASSIFICATION OF B12 SYNTHESIS PATHWAY. THE RANGE WAS SET A PERCENTAGE OF THE FULL SYNTHESIS PATHWAY, WHICH CONTAINS 20 ENZYMES. THE HIGHEST ACCURACY WAS FOUND AT 70% OF THE PATHWAY, TRANSLATING TO SPECIES HAVING MORE THAN 14 OF THE ENZYMES REQUIRED FOR SYNTHESIS.

FIGURE 3.4 – DISTRIBUTION OF B12 SYNTHESIS IN BACTERIAL ORDERS WITH SEQUENCED GENOMES. THE FIGURE SHOWS THE FREQUENCY OF B12 SYNTHESIS IN EACH ORDER. THE HEIGHT OF THE BLUE BARS NEXT TO EACH BACTERIAL ORDER SHOWS THE FREQUENCY IN THE GROUP RANGING FROM 0% (ALL GREY E.G. PASTEURELLALES) TO 100% (ALL BLUE E.G. PROCHLORALAES). THE TREE IS UNROOTED, AND THE COLOURED FIELDS ARE USED TO SEPARATE DIFFERENT PHYLA.

FIGURE 3.5 – DISTRIBUTION OF THE TWO B12 SYNTHESIS PATHWAYS. THE EARLY COBALT INSERTION PATHWAY / ANAEROBIC PATHWAY (ORANGE) AND THE LATE COBALT-INSERTION PATHWAY / AEROBIC PATHWAY (CYAN). THE TREE IS UNROOTED, AND THE COLOURED FIELDS ARE USED TO SEPARATE DIFFERENT PHYLA.

FIGURE 3.6 – PHYLOGENETIC TREE OF 440 SEQUENCED BURKOLDERIALES SPECIES AND ASSIGNMENT OF SPECIES THE EARLY (ORANGE) AND LATE (CYAN) COBALT INSERTION B12 SYNTHESIS PATHWAYS. THE BURKOLDERIA GENUS IS MARKED IN RED. THE SPECIES IN THIS GROUP IS SPLIT APPROXIMATELY EQUAL BETWEEN THE EARLY AND LATE COBALT INSERTION PATHWAY AS SEEN BY THE MIX OF CYAN AND ORANGE BARS.

FIGURE 3.7 – FREQUENCY OF DMB SYNTHESIS AMONG B12-PRODUCERS FROM THE VARIOUS PHYLOGENETIC ORDERS. SPECIES WITH THE BLUB GENE ARE SHOWN IN PURPLE AND SPECIES WITH THE ANAEROBIC (BZA OPERON) DMB SYNTHESIS PATHWAY ARE SHOWN IN GREEN. E.G. 100% OF THE B12-PRODUCERS FROM PROCHLORALAES PRODUCE PSEUDOCOBALAMIN, WHILE 90% OF THE RHIZOBIALES PRODUCE COBALAMIN. THE TREE IS UNROOTED, AND THE COLOURED FIELDS ARE USED TO SEPARATE DIFFERENT PHYLA.
FIGURE 3.8 – FREQUENCY OF B12 SYNTHESIS IN BACTERIA IN SELECTED ENVIRONMENTS. FRACTION WITH THE EARLY COBALT INSERTION PATHWAY IS SHOWN IN ORANGE AND THE LATE COBALT INSERTION PATHWAY IS SHOWN IN CYAN. THE AVERAGE ACROSS ALL SPECIES, REGARDLESS OF ENVIRONMENT IS MARKED BY THE DOTTED LINE. THE NUMBER OF SPECIES ANALYSED FROM EACH ENVIRONMENT IS SHOWN IN SQUARE BRACKETS. E.G. OF THE 103 SPECIES FROM THE RHIZOPLANE, 44% HAVE THE EARLY COBALT INSERTION PATHWAY, WHILE 35% HAVE THE LATE INSERTION SYNTHESIS PATHWAY. THE REMAINING FRACTION ARE NOT ABLE TO SYNTHESISE B12.

FIGURE 3.9 – FRACTION OF SPECIES PREDICTED TO PRODUCE PSEUDOCOBALAMIN (GREY) AND COBALAMIN (GREEN) IN SELECTED ENVIRONMENTS. THE AVERAGE FREQUENCY OF B12 SYNTHESIS ACROSS ALL SPECIES IS MARKED BY THE DOTTED LINE. THE NUMBER OF SPECIES ANALYSED FROM EACH ENVIRONMENT IS DISPLAYED IN SQUARE BRACKETS.

FIGURE 3.10 - VENN DIAGRAM QUANTIFYING B12-DEPENDENT ISOFORMS (METH AND RNR II) AND B12-INDEPENDENT ISOFORMS (METE AND RNR I) OF METHIONINE SYNTHASE AND RIBONUCLEOTIDE REDUCTASE IN BACTERIA. THE NUMBER REPRESENTS THE NUMBER OF BACTERIAL SPECIES WITH EACH OF THE TRAITS, E.G. 3216 BACTERIA OUT OF 8425 HAVE METH, BUT NOT METE.

FIGURE 3.11 – FREQUENCY OF B12 DEPENDENCE IN BACTERIAL ORDERS. B12 AUXOTROPHY WAS ASSIGNED TO SPECIES MISSING THE B12 SYNTHESIS CAPABILITIES AND AT THE SAME TIME-HAVING AN OBLIGATE REQUIREMENT FOR B12. THE HEIGHT OF THE RED BAR NEXT TO EACH GROUP REPRESENTS THE FRACTION OF SPECIES THAT DEPEND ON EXOGONOUS B12, E.G. 86% OF MYXOCOCCALES.

FIGURE 3.12 - FRACTION OF SPECIES THAT ARE CAPABLE OF B12 SYNTHESIS DISTRIBUTION INTO BINS BASED ON THE NUMBER OF PREDICTED PROTEINS EACH SPECIES HAS. THE NUMBER OF SPECIES IN EACH BIN IS SHOWN IN BRACKETS. B12 SYNTHESIS IS FOUND 20% OF SPECIES WITH FEWER THAN 2000 PREDICTED PROTEINS, WHILE IT IS FOUND IN 83% OF SPECIES WITH 9-10K PREDICTED PROTEINS.

FIGURE 3.13 – FREQUENCY OF B12 AUXOTROPHY IN SELECTED ENVIRONMENTS. THE AVERAGE ACROSS ALL SPECIES IS MARKED BY THE DOTTED LINE. THE NUMBER OF SPECIES ANALYSED FROM EACH ENVIRONMENT IS DISPLAYED IN SQUARE BRACKETS. E.G. OF THE 222 FRESH WATER BACTERIA STUDIED, 22% WERE FOUND TO BE B12-AUXOTROPHS.

FIGURE 3.14 – HYPOTHESIS OF BI-DIRECTIONAL B VITAMIN EXCHANGE BETWEEN THE MARINE DIATOM P. MULTISERIES AND BACTERIA FROM THE GENERA SULFITOBACTER AND ROSEOBACTER. IT IS HYPOTHESISED THAT OCEAN ALGAE SUCH AS P. MULTISERIES AND SULFITOBACTER/ROSEOBACTER SPECIES FORM
MUTUALISTIC RELATIONSHIPS WHERE B3 AND B7 ARE EXCHANGED FOR B12. IT IS LIKELY THAT OTHER METABOLITES ARE ALSO EXCHANGED.

FIGURE 3.15 – EXPERIMENTAL VALIDATION OF TWO-WAY B VITAMIN EXCHANGE BETWEEN THE MARINE ALGA OSTREOCOCUS TAURI AND THE MARINE BACTERIUM DINOROSEOBACTER SHIBAE. A) GROWTH ASSAY IN MICROTITER PLATE TO VALIDATE VITAMIN AUXOTROPHIES FOR OSTREOCOCUS TAURI AND DINOROSEOBACTER SHIBAE. BACTERIAL CULTURES WERE ALSO SUPPLEMENTED WITH 1% GLUCOSE. B) GROWTH OF O. TAURI IN AXenic (GREY) AND CO-CULTURES WITH D. SHIBAE (WHITE) WITH AND WITHOUT VARIOUS B VITAMINS. C) GROWTH OF D. SHIBAE IN AXenic (GREY) AND CO-CULTURES WITH O. TAURI WITH AND WITHOUT VARIOUS B VITAMINS. FIGURES FROM COOPER ET AL, (IN PREPARATION)

FIGURE 4.1 – AVERAGE QUALITY SCORE FOR BASE CALLING ON EACH POSITION ACROSS THE 75 BP LENGTH READS FROM THE RNA-SEQ DATASET. THE QUALITY SCORES WERE CALCULATED USING FASTQC (ANDREWS, 2016)

FIGURE 4.2A) HISTOGRAM OF THE LENGTH OF THE CONTIGS IN THE FINAL ASSEMBLY FOR L. ROSTRATA. FIGURE 4.2B) HISTOGRAM OF TRANSCRIPT LENGTH IN V. CARTERI (PURPLE) AND C. REINHARDTII (ORANGE). THE MAJORITY OF TRANSCRIPTS IN THE V. CARTERI AND C. REINHARDTII TRANSCRIPTOMES ARE LONGER THAN 1500 NT, WHILE THE OPPOSITE IS TRUE FOR THE ASSEMBLY FOR L. ROSTRATE.

FIGURE 4.3 - ACCUMULATED COUNTS FOR EACH GO-TERM. LEVEL 2 OF THE GO TAXONOMY WAS USED AND TERMS AT LOWER LEVELS WERE SUMMED TO THEIR LEVEL 2 TERM. THE COUNTS ARE BASED ON GO ANNOTATION OF 16,206 CONTIGS WITH A TOTAL OF 103,082 GO TERMS.

FIGURE 4.4 – FRACTION OF THE READS ASSIGNED TO EACH SPECIES. EACH READ WAS ASSIGNED TO A SPECIES WITH A PUBLIC GENOME BASED ON HOMOLOGY USING BLASTX. THE FIGURE SHOWS THAT 43.38% OF THE READS FROM THE EXPERIMENT MAP TO SEQUENCES WITH A CLOSE HOMOLOG IN C. REINHARDTII.

FIGURE 4.5– SUMMARY OF THE GO-TERMS (LEVEL 3) FOR THE DIFFERENTIALLY EXPRESSED CONTIGS. CATEGORIES FOR CONTIGS UP-REGULATED CONTIGS IN CO-CULTURES ARE SHOWN IN BLUE AND DOWN-REGULATED CONTIGS ARE SHOWN IN RED. A SINGLE CONTIG CAN HAVE MULTIPLE GO-TERMS ATTACHED TO IT.
List of Tables

TABLE 2.1 - EXAMPLES OF PUBLISHED GSM MODELS. FOR EACH ORGANISM IS SUMMARIZED THE NUMBER OF GENES ACCOUNTED FOR BY THE MODEL, THE NUMBER METABOLIC REACTIONS AND THE NUMBER OF METABOLITES INVOLVED IN THOSE REACTIONS. 25

TABLE 2.2 - CONDITIONS THAT SUPPORT SYMBIOTIC BEHAVIOUR FOR M. LOTI AND L. ROSTRATA 30

TABLE 2.3 – STRAINS AND GENOME SEQUENCES USED FOR IN THIS STUDY 31

TABLE 2.4 – EXAMPLE OF THE AUTOMATED ANNOTATION PROCESS FOR DRAFT ASSEMBLY. IN THE EXAMPLE, THE M. LOTI GENE MLR1387 HAS BEEN ANNOTATED AS COBYRINIC ACID, A,C-DIAMIDE SYNTHASE. BASED ON THIS ANNOTATION, THE EC 6.3.5.9 IS IDENTIFIED. THE EC NUMBER IS THEN CORRELATED TO METACYC ID R341-RXN. THIS REVEALS THAT MLR1387 IS PART OF THE B12 SYNTHESIS PATHWAY AND IS RESPONSIBLE FOR THE REACTION SHOWN. 33

TABLE 2.5 - THE BIOMASS OBJECTIVE FUNCTION USED FOR THE M. LOTI AND S. MELILOTI MODELS. THE FUNCTION IS USED TO SIMULATE CONSUMPTION OF METABOLITES AS PART OF GROWTH. THE NUMBER NEXT TO EACH METABOLITE SHOWS THE AMOUNT CONSUMED PER UNIT OF BIOMASS CREATED, E.G. 0.5137 G OF ALANINE IS CONSUMED PER 1 G OF BIOMASS CREATED. 36

TABLE 2.6 – ESTIMATION OF B12 CONTENT PER UNIT OF BIOMASS IN M. LOTI 37

TABLE 2.7 – EXAMPLES OF REACTIONS FROM THE INITIAL DRAFT THAT WERE REMOVED AS PART OF THE CURATION PROCESS BECAUSE THEY DID NOT HAVE DEFINED SUBSTRATES OR PRODUCTS AND THEREFORE COULD NOT BE USED FOR THE FBA. REACTIONS LIKE THESE WERE IDENTIFIED THROUGH SYSTEMATIC INSPECTION OF EACH REACTION IN THE INITIAL MODEL 39

TABLE 2.8 – REACTIONS WHERE REVERSIBILITY WAS CHANGED TO REMOVE CYCLES VIOLATING THERMODYNAMICS. 41

TABLE 2.9 – REACTIONS ADDED TO THE MODEL. FOR ALL BUT ONE REACTION, IT WAS POSSIBLE TO USE GENES FROM OTHER ORGANISMS COMBINED HOMOLOGY
SEARCH TO FIND UNANNOTATED GENES IN *M. LOTI* RESPONSIBLE FOR THE REACTIONS

TABLE 2.10 – COMPARISON OF THE GSM MODEL FOR *M. LOTI* WITH E. COLI AND S. MELILOTI. TOTAL GENES REFER TO THE NUMBER OF GENES IN THE GENOME. INCLUDED GENES IS THE NUMBER OF GENES USE TO GET THE METABOLIC REACTIONS. METABOLITES AND REACTIONS REFER TO NUMBERS FROM THE MODEL OF THE ORGANISMS.

TABLE 2.11 – COMPARISON OF IN SILICO AND IN VIVO GROWTH SUPPORTING SUBSTRATES. METABOLITES PREDICTED TO SUPPORT EXPERIMENTAL GROWTH OF *M. LOTI*, BUT NOT S. MELILOTI WERE TESTED IN A BIOLOG GROWTH ASSAYS. + INDICATES GROWTH SUPPORT AND – INDICATES NOT SUPPORTING GROWTH. FOR THE FOUR SUBSTRATES TESTED SO FAR AND NO DIFFERENCES WERE FOUND BETWEEN THE TWO ORGANISMS.

TABLE 3.1 – COLLECTION OF SEQUENCES FROM THE LATE COBALT INSERTION / AEROBIC PATHWAY FOR B12 SYNTHESIS.

TABLE 3.2 – COLLECTION OF SEQUENCES FOR THE EARLY COBALT INSERTION / ANAEROBIC PATHWAY FOR B12 SYNTHESIS

TABLE 3.3 – COLLECTION OF SEQUENCES FOR SYNTHESIS OF THE LOWER LIGAND OF COBALAMIN

TABLE 3.4 – COLLECTION OF SEQUENCES FOR COBALAMIN DEPENDENT ENZYMES

TABLE 3.5 – COLLECTION OF SEQUENCES FOR COBALAMIN INDEPENDENT ISOFORMS OF THE ENZYMES IN TABLE 3.4

TABLE 3.6 – COLLECTION OF SEQUENCES FOR BIOTIN SYNTHESIS

TABLE 3.7 – COLLECTION OF SEQUENCES FOR NIACIN SYNTHESIS

TABLE 3.8 – COMPARISON OF EXPERIMENTALLY DEMONSTRATED AND PREDICTED PHENOTYPES FOR 43 BACTERIAL SPECIES. THE SPECIES WERE COMPUTATION PREDICTION AND EXPERIMENTAL DATA ARE IN ALIGNMENT ARE ASSIGNED A “+”, WHILE THE SPECIES WHERE THIS IS NOT THE CASE ARE ASSIGNED A “−”. FROM THE COMPARISON IT IS FOUND THAT IT IS POSSIBLE TO CORRECTLY ASSIGN PHENOTYPES IN 41 OUT OF 43 SPECIES GIVING THE METHOD A RAND ACCURACY OF 0.953.

TABLE 3.9 – DEFINITIONS OF VARIOUS CATEGORIES OF B12 METABOLISM FOUND IN BACTERIA AND REFERENCED IN THIS STUDY.

TABLE 3.10 – COLLECTION OF STUDIES OF ALGAE AND BACTERIAL GENERA FROM THEIR PHYCOSPHERE FROM VARIOUS MARINE ENVIRONMENTS. FROM EACH STUDY, THE IDENTIFIED SPECIES AND ALGAE WERE EXTRACTED.

TABLE 3.11 – TABLE OF SEQUENCED SPECIES FROM THE GENERA ROSEOBACTER AND SULFITOBACTER. THE TABLE SHOWS THE PRESENCE (√) OR ABSENCE (−) OF THE SYNTHESIS PATHWAY FOR VITAMIN B12, DMB, VITAMIN B7 AND VITAMIN B3

TABLE 4.2– NUMBER OF RAW READS FROM EACH SAMPLE FROM THE RNA SEQUENCING STEP

TABLE 4.3 – TOTAL NUMBER OF READS FROM EACH CO-CULTURE SAMPLE AND THE NUMBER OF READS ALIGNING TO THE M. LOTI TRANSCRIPTOME FROM THOSE SAMPLES. THE AVERAGE M. LOTI FRACTION OF READS IS 0.06% ACROSS THE SAMPLES.

TABLE 4.4 – READS FROM EACH SAMPLE AFTER IN SILICO REMOVAL OF RRNA AND TRIMMOMATIC FILTERING. THE REMOVAL OF RRNA SEQUENCES WAS AFTER ALIGNING READS TO DATABASES OF RRNA SEQUENCES. THE TRIMMOMATIC FILTERING STEP REMOVED LOW QUALITY READS AND ADAPTER SEQUENCES.

TABLE 4.5 – SIZE AND QUALITY OF THE THREE ASSEMBLIES MADE, THE MERGED ASSEMBLY AND THE FINAL ASSEMBLY. THE FINAL ASSEMBLY IS BASED ON A SUBSET OF THE MERGED ASSEMBLY, SELECTED BASED ON TRANSRATE SCORES.

TABLE 4.6 – OVERVIEW OF ANNOTATION RESULTS. ANNOTATIONS OF CONTIGS WERE MADE USING DIAMOND TO IDENTIFY RELATED SEQUENCES FROM NCBI NR. FROM THE NCBI NR RESULTS, GO AND KEGG CATEGORIES WERE ASSIGNED. CRB-BLAST WAS USED TO IDENTIFY HOMOLOGS FROM THE C. REINHARDTII TRANSCRIPTOME. EXAMPLES ARE PROVIDED IN THE SECOND PART OF THE TABLE.
Abbreviations

B12 – Vitamin B12
B3 – Vitamin B3 (Biotin)
B7 – Vitamin B7 (Niacin)
BHQ – Black queen hypothesis
BOF – Biomass Objective Function
DMB – 5,6-dimethylbenzimidazole
DOM – Dissolved organic matter
FBA – Flux Balance analysis
GSM model – Genome scale metabolic model
HABs - Harmful algal blooms
HGM – Human Gut microbiome
IAA – Indole-3-acetic acid, an auxin class plant hormone
METE/MetE – B12 independent methionine synthase
METH/MetH – B12 dependent methionine synthase
OGs – Orthologous Groups of proteins
RIN – RNA integrity number
TPM – Transcripts per million reads
1 Introduction

1.1 Microbial communities

Microbial communities are found in all environments across the planet and play a central role in aspects such as marine biogeochemistry and for the health of an individual human. They vary in complexity from a few species in an extreme environment such as hot springs, to hundreds of species in the human gut microbiome. Competition for resources such as nitrogen or phosphorus has led to the development of both adversary and symbiotic relationships in these community structures. While the existence of microbial communities has long been known, only through recent technological advances has it become possible to study these important biological structures from a metabolic, genetic and molecular point of view.

The ocean microbiome and human gut microbiome (HGM) are the two most studied microbiomes both in terms of the number of samples and the amount of sequence data. Sunagawa et al, (2015) used a functional analysis approach to identify the overlap between these two microbiomes by comparing the identified Orthologous Groups of proteins (OGs) (Galperin et al, 2014) from the Tara Ocean Project (Sunagawa et al, 2015; Lima-Mendez et al, 2015; Malviya et al, 2016) with the gene catalogue from the human gut established by Qin et al, (2010). Despite the physiological difference between the environments, such as oxygen availability and nutrient concentrations, 73% of the OGs found in the ocean microbiome was also identified as part of the core gene set from the HGM.

1.1.1 Ocean Microbiome

In the oceans, microorganisms play an essential role in biogeochemical process such as carbon fixation, and nutrient cycling. Phytoplankton (cyanobacteria and eukaryotic microalgae) are responsible for approximately half of the net primary
production, which is defined as the gross amount of carbon fixation minus respiration of the carbon fixing organisms (Whitman et al., 1998; Field et al., 1998). The diversity of the ocean microbiome is vast and oceanographers have studied this microbial community for more than four decades ever since it became clear that microbes consume the majority of energy that is being produced in the oceans (Moran, 2015). The study of the ocean microbiome has attracted increased interest as the changes in climate are thought to affect the composition and productivity of this microbial community. As the global temperature of the oceans rise, it leads to changes in currents and mixing dynamics of the oceans, which changes availability of growth limiting nutrients such as iron and nitrogen. As this availability changes, the productivity of the oceans, in terms of carbon fixation, could change and affect the global climate (Behrenfeld et al., 2006).

The threat of changes to ocean productivity has led to studies attempting to increase the knowledge of the mechanisms driving primary productivity, including studies of taxonomical diversity and nutrient availability. One of the biggest of these studies is the TARA ocean studies, (Moran, 2015; Sunagawa et al., 2015), where a research vessel sailed across the globe collecting samples from 68 sites across the Pacific, Atlantic, Southern, Mediterranean, Red Sea and Indian ocean over 3 years. Using Illumina sequencing and sequence mapping, they identified ~40 million representative genes with 59% thought to stem from bacteria, 3.3% from eukaryotes, 27.7% from unknown organisms and the remaining fraction from vira and archaea (Sunagawa et al., 2015). This shows that bacteria are main contributor to the gene diversity of the oceans. Using the sequence data from the oceans samples, the taxonomical landscape of the bacteria was also investigated, and ~40,000 operational taxonomical units (OTUs) were identified, a feature commonly used as a proxy for microbial species (Schloss & Handelsman, 2005; Blaxter et al., 2005). Identification of OTUs are based on clustering 16S rRNA (in the case of bacteria) sequences with high degree of sequence similarity, typically 95%-99%. The threshold reflects the taxonomical precession required, e.g. >97% sequence similarity when trying to group sequences to species level, and >95% when grouping sequences to genus level (Schloss & Handelsman, 2005). This analysis showed that Alpha- and Gamma-proteobacteria are the two most dominant groups in terms of abundance and
make up 2/3 of the ocean microbiome. *Delta-proteobacteria, Cyanobacteria* and *Bacteroidetes* make up the majority of the remaining microbiome (Sunagawa *et al*, 2015). The sequence data from the ocean samples was also subjected to a functional analysis using the eggNOG platform (Powell *et al*, 2012; Huerta-Cepas *et al*, 2016) to identify OGs. 39,246 OGs were identified across all samples, but only 5,755 OGs (14.5%) could be identified in all environmental samples, suggesting a high degree of variation in the functionality of the ocean microbiome in different locations at different times (Sunagawa *et al*, 2015). Interestingly, only 60% of the 5,755 OGs have a known function indicating the need for further studies into the functionality of the ocean prokaryote microbiome.

Using the same samples from the TARA ocean project, a characterisation of the eukaryotic, primarily algae, variation in the oceans was also made. In a study of diatom abundance and taxonomical richness, 4,748 OTUs were identified and 58% of them were mapped to a known genus (Malviya *et al*, 2016). The most abundant diatom genera across all the samples were *Chaetoceros*, which made up 23.1% of the mapped sequences. *Chaetoceros* has attracted a commercial interests as a nutrient in fisheries and potential source for biofuel due to its high lipid content (Ifuku *et al*, 2015; Gu *et al*, 2016). 15.5% of sequences mapped to the genus *Fragilariopsis*, which is the major species in the polar regions and known to produce multiple anti-freeze proteins (Uhlig *et al*, 2011). This was followed by *Thalassiosira* (13.7%) *Corethron* (11%), *Leptocylindrus* (10.1%), *Actinocyclus* (8.7%), *Pseudo-nitzschia* (4.4%), and *Proboscia* (3.9%) (Malviya *et al*, 2016).

### 1.1.2 Human Gut Microbiome

The human gut microbiome (HGM) is likely the most studied microbiome in terms of data produced, samples taken, and research papers published. Recognition of the facts that HGM: 1) contains as many bacterial cells as there are human cells in an individual (Sender *et al*, 2016); 2) has a gene set 150x larger than the human genome (Qin *et al*, 2010); and 3) has a major impact on human health (Cho & Blaser, 2012) has led to large interest in this research area. Two large scale projects, the Human Microbiome Project (Turnbaugh *et al*, 2007) and
European equivalent MetaHIT (Ehrlich, 2011), have carried out sequencing efforts equivalent to the Human Genome Project in order to determine the composition and functionality of what has been dubbed ‘the other genome’ (Zhao, 2010; Brüls & Weissenbach, 2011). Correlations between the particular HGM composition and a wide variety diseases such as diabetes (Qin et al, 2012; Forslund et al, 2015), cardiovascular disorders (Wang et al, 2011), psoriasis (Gao et al, 2008), and more, have been identified. The studies of diseases in relation to the HGM are largely based on comparing the microbiome composition of healthy and sick individuals and identifying differences. While these sequence-based studies allows for detection of correlation between changes in the microbiome and certain diseases, they do not provide evidence for causality, and further studies are required before the knowledge can be applied in a clinical setting (Cho & Blaser, 2012).

Considering the number of bacterial species present in the gut microbiome, the potential for variation between humans is almost unlimited. It was expected that the microbiome species composition was correlated with factors such as age, body mass index, gender or geographical location (e.g. difference between continents). However based on more than 400 metagenomes and using principle component analysis, three distinct clusters (dubbed enterotypes) that are independent of the mentioned factors were identified (Arumugam et al, 2011). The three enterotypes are characterised by the abundance of either one of the following bacterial genera: Bacteroides (enterotype 1), Prevotella (enterotype 2) and Ruminococcus (enterotype 3). In addition to the species identification, Arumugam et al, (2011) also performed a functional analysis, quantifying abundance of OGs. While vitamin synthesis pathways could be found in all three enterotypes, enterotype 1 was enriched in biotin, riboflavin, pantothenate and ascorbate synthesis pathways genes, while enterotype 2 was enriched in thiamine and folate synthesis genes. The impact of these functional differences on human health is still unclear, but it is hypothesised that availability of limiting vitamins can shape a microbial community and vitamin supplements could be used to alter the gut microbiome and thus improve human health.
1.2 Microalgae

1.2.1 Microalgae and their global implications

Microalgae in the marine environment are thought to be responsible for up to 50% of the global carbon fixation through photosynthetic processes and are thus of high importance as primary producers underpinning more complex lifeforms (Falkowski, 1994). ‘Microalgae’ is a broad term for unicellular photosynthetic organisms. The species in this group are not strictly related from an evolutionary point of view as the term covers diverse groups, such as diatoms, dinoflagellates, the green algae (Chlorophytes and Charophytes) and more. The phylogenetic difference between species of algae can be vast, with the main common denominator between members being that they are eukaryotes and contain plastids, which originate from an endosymbiotic event involving cyanobacteria (Cooper & Smith, 2015) (Figure 1.1).

Figure 1.1 – Overview of algal phylogeny from ancient ancestral organisms to current species. Figure from Cooper & Smith, (2015)
While algae through their role as primary producers support all secondary life forms, they can also cause major environmental havoc. Among the eukaryotic algae, species of diatoms and dinoflagellates are known to produce harmful algal blooms (HAB) in oceans, such as the annual red tide blooms along the west coast of the US. There the toxins domoic acid from the *Pseudo-nitzschia* (diatom) genus and saxitoxin from genus *Alexandrium* (dinoflagellate) cause paralytic shellfish poisoning (Horner et al, 1997). Most studies of algae blooms and the attempts to understand why they form, have been focused on availability of iron, phosphorus and nitrogen and how a sudden influx of these inorganic nutrients might lead to an algae bloom. However, a more recent study has shown that a large part of these harmful algae are dependent on Vitamin B1 and B12 and that availability of these vitamins might impact blooming (Tang et al, 2010).

1.2.2 Algae-bacterial interactions

The oceans are generally dilute both in terms of nutrients such as iron, nitrogen and phosphorus and while algae can fix CO₂ and cover their energy need through photosynthesis, many rely on other microorganisms for micronutrients. The oceans are estimated to contain ~10⁶ bacterial/mL (Whitman et al, 1998) and ~10³ eukaryotic algae/mL (Andersen, 1992), and surrounding the algae is a thin layer of fluid called the phycosphere (Bell & Mitchell, 1972). In this microenvironment algae secrete dissolved organic matter (DOM) such as simple sugars, amino acids, organic acids, complex polysaccharides, and lipids (Sharp, 1977; Hedges *et al*, 2001; Amin *et al*, 2015). In the phycosphere the amount of DOM decreases exponentially as the distance from the algae increases (Seymour *et al*, 2010), which leads to bacteria accumulating in close proximity to the algae due to the chemotaxis by bacteria (Stocker & Seymour, 2012). While the bacteria in the phycosphere are dependent on the DOM provided by the algae, they affect the algae and phycosphere by providing micronutrients and making sparse nutrients more bioavailable.

As the iron concentration in oceans is low, it can often be a growth limiting nutrient for algae (Bertrand *et al*, 2007, 2015). Amin *et al*, (2009) found that members of the *Marinobacter* genus, which are known to form communities with
dinoflagellates such as *Scrippsiella trochoidea*, produce an iron binding siderophore called vibrioferrin, which has an average affinity for iron, but a high photolysis rate. The presence of this siderophore results in 20x increase in iron uptake by the algae and a 70% increased uptake by the bacteria. It is therefore a greater advantage for the algae than the bacteria, despite it being produced by the bacteria. This difference is thought to be due to the absorbance of light by the algae and the high photolysis rate. It was also found that *S. trochoidea* provide DOM to the *Marinobacter*, but the exact metabolite was not identified. Surprisingly, vibrioferrin was also shown to facilitate uptake of boron, which is essential in some phytoplankton and has been hypothesised to have multiple roles in interspecies communication (Lewin & Chen, 1976; Amin et al, 2007).

Using a model system of the diatom *Thalassiosira pseudonana* and the *alpaphroteobacteria* *Roseobacter pomeroyi*, a study of the metabolites exchanged between these two representative species was undertaken by Durham et al, (2014). The authors carried out a transcriptome analysis using RNA-seq of co-cultures of the two organisms, and axenic cultures, which enabled identification of differentially expressed transcripts between the two experimental conditions. A transcript encoding a protein involved in transport and metabolism of 2,3-dihydroxypropane-1-sulfonate in *R. pomeroyi* was found to be upregulated 374-fold in co-cultures, which suggest that this metabolite is used by the algae to attract beneficial bacteria. It was demonstrated that the B12-requirement of *T. pseunonana* can be met by *R. pomeroyi*, which has been shown to carry the genes required for B12 synthesis (Sañudo-Wilhelmy et al, 2014).

A similar study of the diatom *Pseudo-nitzschia multiseries* and the *alpaphroteobacteria* *Sulfitobacter SA11* was performed with the aim of identifying metabolites exchanged by the two species (Amin et al, 2015). The authors demonstrated that the bacteria enhance growth of the diatom and using a transcriptomics approach, they identified an upregulation in the indole-3-acetic acid (IAA) synthesis pathway. IAA is a known plant hormone of the class called auxins and stimulates growth. The effect of auxin on algae in general is still poorly understood, but it was demonstrated to increase growth of *P. multiseries* and to be involved in cell division regulation (Lau et al, 2009; Amin et al, 2015). It was also shown that the auxin is unlikely to be responsible for the entire enhanced
growth as an axenic culture of *P. multiseriess* supplemented with IAA has a lower growth rate compared to *P. multiseriess* in co-culture with *Sulfitobacter SA11*. When studying the algae side of the interaction, it was discovered that the algae supply the bacterium with DOM in the form of the sulphur containing dimethylsulfoiniopropionate as well as tryptophan. Tryptophan is particularly interesting as it is a precursor for IAA, and thus demonstrates that the interaction between the two organisms is of a complex nature.

Nitrogen is an essential macronutrient for algae and as they are unable to fix N\(_2\), they rely on other microorganisms for this resource. A study of the marine cyanobacteria species *Richelia intracellularis* and *Calothrix rhizosoleniae* demonstrates that these organisms can form communities with various diatoms (Foster *et al*, 2011). Using a type of mass spectrometry called nanoSIMS, the N\(_2\) fixation rates of the cyanobacteria were estimated in both co-cultures and in axenic cultures and found to be 171–420x higher when the cyanobacteria were co-cultured with algae. The experiment also showed that the 97% of the fixed nitrogen ultimately ended up in the algae.

In the examples above, the interaction between the algae and bacteria is beneficial or at least neutral for both parties. The widespread marine algae, *Emiliana huxleyi* can form communities with the alpha-proteobacterium *Phaeobacter gallaciensis* (Seyedsayamdost *et al*, 2011a, 2011b) and at the onset of the interaction, *E. huxleyi* provides an organosulfur that the bacteria can use as an energy source, while *P. gallaciensis* supports the algae by providing a growth enhancing auxin and secreting antibiotics. As the co-cultures mature the algae begins releasing p-coumaric acid, which stems from the lignification of the algae cell wall (Martone *et al*, 2009). This prompts the bacteria to begin synthesising and secreting a class of algicides dubbed roseobacticides, which leads to algae death, and then the interaction switches from mutualistic to parasitic.

The examples above demonstrate that algae and bacteria can form mutualistic relations in nature, but it is also possible to construct synthetic communities of algae and bacteria. Using the freshwater green alga *Lobomonas rostrata*, which is related to the model alga *Chlamydomonas reinhardtii*, and the soil alpha-proteobacterium *Mesorhizobium loti*, it is possible to construct a stable and
mutualistic community (Kazamia et al., 2012b; Grant et al., 2014). L. rostrata can cover its need for organic nutrients through photosynthesis, except for vitamin B12 (cobalamin), which it requires from an exogenous source, while M. loti produces B12, but requires fixed carbon as an energy source. When the two organisms are co-cultured in a media with neither DOM nor B12, they both grow, forming a community with ~30 bacteria per algal cell. Interestingly, this ratio of algae to bacteria appears to be independent of the initial concentration of the inoculum (Kazamia et al., 2012b), and indicates that each is regulating the growth of the other species (Grant et al., 2014).

The paragraphs above have provided examples of both natural and synthetic communities of algae and bacteria with the majority of these communities being mutualistic. In Figure 1.2 below, the metabolites exchanged between the two types of organism have been summarised, however it should be noted that all of these groups of metabolites might not be found to be exchanged in a specific algae-bacterial community. In the studies discussed, it is mentioned that these interactions are complex and that it is unlikely that all the mechanisms and metabolites involved in the action have been uncovered. As shown, the interaction between algae and bacteria in the phycosphere is complex and it has been proposed that a specific interaction might have started with exchange of a single compound and then evolved to include more metabolites (van Tol et al., 2016).
1.2.2.1 A theory on the evolution of communities

A recent theory of how communities develop and evolve is “The Black Queen Hypothesis” (BQH), which is based on a study of the widespread marine cyanobacterial genus *Prochlorococcus* (Morris *et al.*, 2012). *Prochlorococcus* has one of the smallest genomes among cyanobacteria with only ~1700 genes (Rocap *et al.*, 2003) compared with the closely related *Synechococcus WH8102*, which has ~50% more genes (Palenik *et al.*, 2003). This apparent loss of genes in *Prochlorococcus* includes the loss of an arsenal of anti-oxidative stress genes particularly *katG*, which encodes catalase-peroxidase that protects against H$_2$O$_2$ damage (Morris *et al.*, 2008; Scanlan *et al.*, 2009). As *katG* requires iron to function, *Prochlorococcus* MED4 has a lower metabolic requirement for iron, which is a growth advantage as iron is often growth limiting in oceans (Bertrand *et al.*, 2007). Further studies showed that when the *Prochlorococcus* species is co-cultured with the heterotrophic *Alteromonas EZ55*, which does have the *katG* gene, *Prochlorococcus* is able to grow in seawater, because H$_2$O$_2$ is reduced to...
non-toxic levels (Morris et al, 2011). The above observations led to the formulation of the BQH, which described the evolution of communities using a game theory approach (Morris et al, 2012). The BQH proposes that the small genomes observed are results of adaptive genome reduction that occurs when a common pool of a particular ‘good’ (such as a nutrient or detoxification in the case of Prochlorococcus/Alteromonas) is available. The benefitting organisms stop contributing to this pool, but instead rely on ‘helper’ organisms for this ‘good’. The ‘beneficiary’ gains an advantage, since it doesn’t need to spend energy on making common ‘good’, but at the same time it is now dependent on ‘helpers’ for this particular ‘good’. This puts evolutionary pressure on the ‘beneficiary’ to maintain association with members of the ‘helper’ group. The theory explains how dependencies between organisms can occur as a result of interactions and as a way of dividing labour. A complementary theory to the BQH dubbed ‘forager to farmer’ has been proposed based on observations of algae and bacteria and evolution of mutualism (Kazamia et al, 2016). C. reinhardtii is a B12-independent alga but will ‘forage’ and utilise this metabolite when it is available. However, when supplied with an abundance of B12 over a longer period, an evolved strain of C. reinhardtii arose that was B12-dependent (Helliwell et al, 2014a). If this occurred throughout the algal kingdom more generally, it would explain the widespread distribution of B12 auxotrophy, which is found in approximately half of all algal species (Croft et al, 2005). After becoming B12-dependent, the algae would need an exogenous source of B12 and might start ‘farming’ bacteria by providing DOM in its phycosphere.

1.3 Vitamin B12

1.3.1 Vitamin B12 as an enzymatic co-factor

B12 is one of the largest and most complex molecules found in nature. It is made up of a corrin ring with a cobalt ion at the centre, an upper axial ligand and a nucleotide loop that supports the lower axial ligand. Interest in this micronutrient was prompted by Minot & Murphy (1926), who discovered that liver extract could be used to treat pernicious anaemia. The molecule itself was isolated in 1948 (Rickes et al, 1948; Smith, 1948) and using x-ray crystallography Hodgkin et al, (1955) managed to identify the structure of this complex B12 (Figure 1.3)
B12 acts as a co-factor for three classes of enzymes, isomerases, methyltransferases and reductive dehalogenases (Banerjee & Ragsdale, 2003). In mammals, two particular B12-utilising enzymes are essential for maintaining a proper health: methylmalonyl-CoA mutase and methionine synthase (Marsh, 1999; Banerjee & Ragsdale, 2003). Methylmalonyl-CoA mutase, which is encoded by the MUT gene, is an isomerase that facilitates conversion of methylmalonyl-CoA to succinyl-CoA, which a central aspect of odd-chain fatty acid metabolism. Methionine synthase is encoded by the METH gene and is responsible for methylating homocysteine and converting it into methionine. Homocysteine is toxic in high concentrations and when B12 is absent, homocysteine is likely to accumulate, which can lead to pernicious anaemia (Marsh, 1999). In lower amounts it is correlated with ischaemic heart disease. In addition to these two enzymes, a B12-utilizing isomerase, ribonucleotide reductase (type II) is found in bacteria. This enzyme catalyses conversion of ribonucleotides to deoxy-ribonucleotides that are essential for DNA-replication and repair (Gleason & Olszewski, 2002; Dawes et al, 2003; Warner et al, 2007). As DNA-replication is an essential process for any living organisms attempting to replicate, and RNR is the only group of enzymes capable of converting ribonucleotides into deoxy-ribonucleotides, this enzyme is thought to essential for all bacteria (Torrents, 2014), although it should be pointed out that there are B12-independent forms of RNR (see below). The third class of B12-dependent enzymes, the reductive dehalogenases, is found in anaerobic bacteria and facilitates dehalogenation, which is the process of removing halogens such as
chlorine and substituting them for either a hydrogen or in cases of alkane/alkenes for a new carbon-carbon bond (Mohn & Tiedje, 1992). As several halogens, including those from pesticides, are environmentally damaging, dehalogenation plays an important role in bioremediation (Mohn & Tiedje, 1992). Reductive halogenases are membrane bound enzymes that are highly oxygen-sensitive, which have made detailed studies challenging (Payne et al, 2014). By using a non-oxygen-sensitive reductive halogenase from the bacteria *Nitratireductor pacificus*, it was shown that the cobalt from cobalamin is directly involved in the dehalogenation process and that this process is thus unlikely to occur without a cobalt ion present (Payne et al, 2014).

Both methionine synthase and ribonucleotide reductase exist in a B12-dependent form and in a B12-independent form (Figure 1.4). In bacteria, the methionine synthase isoforms are encoded by the *metE* (B12-independent) and *metH* (B12-dependent) genes. Mammals and higher eukaryotes only have the B12-dependent form of this enzyme, while bacteria and algal species can have just *metE* or *metH*, or both of them. Species that have both genes, such as the green alga *C. reinhardtii* or the diatom *Phaeodactylum tricornutum*, have been shown to utilise the B12-dependent isoform when B12 is available and the B12-independent form when B12 is scarce (Warner et al, 2007; Helliwell et al, 2014b). The conversion of ribonucleotides into the essential deoxy-ribonucleotides can be carried out by 3 enzymes, RNR class I (consisting of subclasses Ia and Ib) and class III, which are B12-independent, as well as RNR class II, which requires B12. As with methionine synthases, any bacterial species can possess either one or more of the three classes and having multiple classes enable the species to adapt to changing environments and substrate availabilities. Classes I and III do not need B12 as cofactor, but class I requires oxygen, while class III only functions under anaerobic conditions. Class II can function under both aerobic and anaerobic conditions and having this class therefore enables a bacterium to survive in environments that switch between aerobic and anaerobic states (Torrents, 2014). An example of a bacterium that utilises all three classes is the biofilm forming opportunistic pathogen *Pseudomonas aeruginosa*, which can live both in aerobic environments such as the soil, where it utilises RNR class I and II, and in an anaerobic environment such as the lower layers of a biofilm in the
lungs of humans, where a combination of class II and III are active (Sjöberg & Torrents, 2011; Crespo et al, 2016). In addition, some species have been shown to have genes coding for two different variants of class I RNR. E. coli is such a species, which has been shown to utilise class Ia variant under most conditions, but the class Ib enzyme when iron is limiting (Martin & Imlay, 2011). Together these multiple versions of the ribonucleotide reductase allow bacteria to carry out this function under a wide range of environmental conditions (Torrents, 2014).

While enzymes that utilise B12 as cofactor can be found in all mammals, in some eubacteria (Warren et al, 2002), algae (Croft et al, 2005) and archaea (Thomas & Escalante-Semerena, 2000; Torrents et al, 2002), but in land plants (Smith et al, 2007) and fungi (Suliman et al, 2005), there are no B12-dependent enzymes. Plants and fungi use the B12-independent methionine synthase (MetE) to cover their methionine need, and RNR I to catalyse deoxy-ribonucleotide synthesis. Moreover, B12 cannot be detected in these organisms, which suggests that they do not produce it either.

1.3.2 Vitamin B12 dependence in algae

In the studies described in section 1.2.2, B12 was demonstrated as a nutrient being supplied by bacteria to algae in natural environments. It has been demonstrated that approximately half of eukaryotic algae are dependent on exogenous B12 (Croft et al, 2005) and presently no eukaryotic algae have been
demonstrated to synthesise B12 de novo. As the species of algae that are B12-dependent are not closely related, it is thought that B12-dependence in algae is a trait that has risen multiple times throughout evolution. Using data from sequenced algae, genes for three B12-dependent enzymes were identified, methylmalonyl-CoA mutase, METH, and RNR II, although many species encoded just one of these (Helliwell et al., 2011). By comparing the experimental data and genomic data, B12 auxotrophy in algae appeared to be related to the loss of the B12-independent methionine synthase gene, METE. In the algal species without the METE gene, this essential function is carried out by the B12-dependent enzyme encoded by the METH gene. It is not known what has caused this dependence to evolve, but culturing of the B12-independent alga, C. reinhardtii in excess B12 led to the evolution of a B12 auxotroph in fewer than 500x generations (Helliwell et al., 2014a). Sequencing of the METE gene in this strain revealed that a transposable element was inserted into the gene rendering it functionless. It was also demonstrated that the B12 producing bacteria M. loti was able to provide B12 and rescued growth of the algae. This points towards B12 auxotrophy evolving in algae when algae are in communities where B12 is readily available.

1.4 Other B-vitamins

1.4.1 Vitamin B7 (Biotin)

Biotin was first discovered in 1901 and demonstrated to affect growth of yeast. Biotin is involved in fatty acid synthesis, branched-chain amino acid catabolism, and gluconeogenesis (Wood & Barden, 1977; Campbell & Cronan, 2001; Tong, 2005; Waldrop et al., 2012) and is an essential vitamin in all organisms (Lin et al., 2010). It is a cofactor for three classes of enzymes, carboxylases (e.g. pyruvate carboxylase, which is part of the gluconeogenesis pathway), transcarboxylases (e.g. Transcarboxylase that is central to propionic acid fermentation in bacteria) and decarboxylases enzymes (e.g. methylmalonyl-CoA decarboxylase which is part of succinate fermentation) (Murtif et al., 1985; Bott et al., 1997; Sañudo-Wilhelmy et al., 2014). Biotin is made up of two heterocyclic rings and a valeric acid side chain (Figure 1.5A) and is produced by both eukaryotes, prokaryotes
and archaea. The synthesis pathway branches off the fatty acid synthetic pathway (Lin et al., 2010) and takes different routes in different organisms to arrive at the precursor pimeloyl-CoA (Entchev et al., 2002; Wu et al., 2005). In bacteria, four enzymes, \textit{bioF}, \textit{bioA}, \textit{bioD} and \textit{bioB}, which are organised in an operon, are responsible for converting pimeloyl-CoA into biotin (Figure 1.5B, Marquet et al., 2001). The plant \textit{Arabidopsis thaliana} has genes analogous to \textit{bioF} (BIO4 in \textit{A. thaliana}), and \textit{bioB} (BIO2) with the function of \textit{bioA} and \textit{bioD} being carried out by bi-functional protein called BIO3-BIO1 (Baldet et al., 1997; Pinon et al., 2005; Muralla et al., 2008). In algae, the biotin synthesis pathway has not been studied in detail, but homologous of \textit{bioF}, BIO3-BIO1 and \textit{bioB} can be found in algae such as \textit{C. reinhardtii} and \textit{Thalassiosira pseudonana} (Croft et al., 2006).

While all organisms require biotin, not all are capable of synthesis. Biotin auxotrophy is found in all mammals, which requires indigested biotin in order to maintain proper health (Landenberger et al., 2004). In other eukaryotes such as algae, a previous study found that 13 out of 306 (4.5%) surveyed species were biotin-auxotrophs and this trait is thus considered rare (Croft et al., 2006). In a study focused on algal species causing harmful algal blooms (HABs) it was found that 10 out of 27 investigated species require a supplement of biotin in order to grow (Tang et al., 2010). Based on these two studies it appears that there is an overrepresentation of biotin auxotrophy among the algae causing HABs compared to other algae. The reason for this has not been uncovered, but it suggests that algae causing HABs are required to interact with a biotin-producing algae or bacteria in the oceans.

In a study of marine bacteria Sañudo-Wilhelmy et al., (2014) surveyed 413 bacterial species using a comparative genomics and found that in ~80% of bacteria have homologs to \textit{bioB}, while no match could be found in the remaining ~20% suggesting that these bacteria are biotin-auxotrophs. It should be noted here however that only \textit{bioB} was searched for, not any of the other three enzymes in the synthesis pathway. By including the other enzymes from the biotin synthesis pathway, it will likely reveal that the true fraction of biotin-auxotrophs is even larger. While the above study is based only on sequence information, there are also experimental data demonstrating biotin auxotrophy among oceanic bacteria. The two marine \textit{alpha-proteobacteria} \textit{Roseobacter denitrificans} and
*Roseobacter litoralis* have been demonstrated to require exogenous biotin for growth (Shiba, 1991). The biotin auxotrophy is potentially caused by the absence of *bioD* and *bioB* homologs in the two *roseobacter* species.

In a study of the HGM, 259 bacterial species were studied with a comparative genomics approach using annotated genomes from PubSeed (Overbeek *et al.*, 2005), and each species was classified as either having the biotin synthesis pathway or not (Magnúsdóttir *et al.*, 2015). It was found that ~40% of the studied bacteria were predicted to be capable of biotin synthesis, which suggests that the remaining 60% of the bacteria are dependent on exogenous biotin. If the predicted ratio of biotin-producers/auxotrophs is true, biotin availability could have a strong impact on the composition of microbiome.

![Figure 1.5](image)

**Figure 1.5**—A) Structure of biotin (Hastings *et al.*, 2013) B) Biotin synthesis pathway from the precursor Pimeloyl-CoA to the final form of biotin. Genes from plants/algae are in green, while genes from bacteria in blue.

### 1.4.2 Vitamin B3 (niacin)

Niacin is an enzymatic co-factor for synthesis of nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP). These two molecules are redox co-factors, thought to be essential in all life forms and involved in more than 100 redox-reactions in important pathways such as citric acid cycle, fatty acid synthesis and photosynthesis (Sistare & Haynes, 1985; Nicholls, 2013). In humans, niacin-deficiency is known as pellagra and can cause diarrhoea, dermatitis, and dementia, and be potentially fatal if left untreated (Hegyi *et al.*, 2004; Prakash *et al.*, 2008). Pellagra was first described by Gasper Casa in 1762, who noted that people in Andalusia who had a diet mainly consisting of maize developed sores on their hands and feet. In 1914 it was discovered that it was a result of nutrient deficiency. While maize does contain
niacin, it is bound in a form that makes it unavailable to humans, unless it is boiled in alkaline water such as limewater (Hegyi et al., 2004). In 1922 Goldberger & Tanner (1922) found that amino acid supplementation could improve conditions in Pellagra patients and finally in 1937 it was found that niacin was able to cure the condition. Humans cover the majority of their niacin requirements through uptake from their diet but can synthesise niacin from tryptophan degradation. This process however is limited by available tryptophan as 60 mg of tryptophan converts to only 1mg of niacin (Hegyi et al., 2004).

Niacin can be synthesised either from aspartate or tryptophan, and the conversion from tryptophan can be done via two different routes (Figure 1.6). In the ‘aspartate pathway’, aspartate is converted to iminoaspartate by the gene product of nadB or nadX (Yang et al., 2003). While the two enzymes catalyse the same conversion, they appear to be unrelated based on sequence similarity (Yang et al., 2003). Quinolinate synthase is encoded by nadA and converts iminoaspartate to quinolinate and from there the enzyme encoded by nadC is responsible for the conversion of quinolinate to niacin. The tryptophan pathways are based on tryptophan being degraded to either anthranilate or kynureine. In the anthranilate branch, anthranilate is converted to 3-hydroxyanthranilate using hpaB (Liu et al., 2010). In the kynureine branch, kynureine is converted to 3-hydroxy-L-kynurenine by kynurenine 3-monooxygenase (KMO), which is then converted to 3-hydroxyanthranilate by kynureninase (KYNU). At this point the two tryptophan branches converge and the remaining reactions are identical in both branches. 3-hydroxyanthranilate is converted to 2-amino-3-carboxymuconate semialdehyde by HAAO (Malherbe et al., 1994) and in a non-enzymactic reaction this molecule is converted to quinolinate (Belenky et al., 2007). The final step of the tryptophan pathway is the conversion of quinolinate to niacin catalysed by nicotinate-nucleotide pyrophosphorylase (encoded by nadC) and identical to the aspartate pathway (Hughes et al., 1993; Rongvaux et al., 2003; Kurnasov et al., 2003). Mammals synthesise niacin from tryptophan (Rongvaux et al., 2003), plants use aspartate as a precursor (Noctor et al., 2006) and bacteria have been found to use both (Kurnasov et al., 2003). Niacin synthesis in algae has not been extensively studied, but algal species such as Chlorella vulgaris and Chlorella
pyrenoidosa have been demonstrated to produce this vitamin (Pratt & Johnson, 1964).

As niacin is used by all organisms and produced by most microorganisms, niacin-auxotrophy among microorganisms has received less interest compared with B12 or biotin auxotrophy studies. In bacteria, there are a number of unrelated examples of niacin-auxotrophy resulting from the absence of one or more niacin synthesis genes. One example of this is the opportunistic pathogen *Staphylococcus aureus* from the *firmicute* phylum, which can cause skin infections. In this bacterium the auxotrophy is caused by the absence of the *nadC* gene (Knight, 1937; Bosi *et al*, 2016). It is not clear how niacin-auxotrophy affects the life cycle of the bacterium, but since niacin can be a growth limiting factor, it presents a potential therapeutic target. Another pathogen that is a known niacin-dependent organism is *Shigellae dysenteriae*, which is a gamma-proteobacteria from the same family as *E. coli* (Ahmed *et al*, 1988). *S. dysenteriae* causes infections in the gut, known as dysentery, by secreting shiga toxin, which inhibits protein synthesis (Sandvig & van Deurs, 2000). The niacin source for *S. dysenteriae* has not been demonstrated directly, but as more than ~60% of bacteria in the human gut are predicted to synthesise niacin (Magnúsdóttir *et al*, 2015), the niacin-dependent bacteria might form some form of community with niacin-producing bacteria. The human pathogen *Streptococcus pyogenes*, which causes skin infections and is responsible for >500,000 deaths annually (Cohen-Poradosu & Kasper, 2007) has also been found to require exogenous niacin (Sorci *et al*, 2013). On the skin *S. pyogenes* is a part of community consisting of a large variety of bacteria (Grice & Segre, 2011), which is a potential source for the essential niacin, and niacin availability might play an important role during infections. In the alpha-proteobacteria *R. denitrificans* and *R. litoralis*, found in the marine environment, niacin auxotrophy has been demonstrated experimentally (Shiba, 1991) and is potentially caused by the absence of *nadA* from the aspartate pathway or absence of multiple enzymes from the tryptophan pathways.
Figure 1.6 – Niacin synthesis pathway. Niacin can be synthesised from either aspartate or tryptophan through a series of enzymatic steps. There are two variants of the tryptophan pathway. The enzyme names are mentioned as well as the intermediates in the synthesis pathway.
1.5 Aim and structure of this thesis

My overarching aim for this thesis was to study how algae and bacteria interact, and to improve the understanding of which metabolites and mechanisms that are central to this interaction. My research is largely based on the model system of bacterium *M. loti* and alga *L. rostrata*, where vitamin B12 is provided by the bacterium, and the alga partner provides the community with fixed carbon (Kazamia *et al.*, 2012b; Grant *et al.*, 2014). The thesis is structured into three distinct research chapters each with their own introduction, results and discussion and finished with a discussion chapter placing the findings in a broader context.

In chapter 2, I investigated how metabolic modelling could be applied to study the metabolism of *M. loti*. The aim of this study was to simulate growing *M. loti* using various substrates as energy source to identify substrates that could be provided by *L. rostrata* to *M. loti* in co-cultures of the two organisms.

In chapter 3, I present my findings from a broad study of B12 synthesis, usage and dependence in bacteria. This chapter was motivated by the lack of knowledge about which bacteria are capable of B12 synthesis. Using comparative genomics techniques, the presence of more than 50 genes in more than 8000 bacterial species was analysed to provide an overview of B12 synthesis and utilisation in bacteria.

The study presented in chapter 4 was motivated by a wish to identify genes that play a key role in the mutualistic community of *M. loti* and *L. rostrata*. I set out to identify genes that were differentially expressed in the two organisms when comparing an axenic state to a co-culture of the two organisms using transcriptomics.
1.6 Publications

In addition to the research presented in this thesis, I co-authored two published studies, one that has been accepted for publication, as detailed below, and included as Appendices.

In a paper on cobalamin and pseudo-cobalamin in cyanobacteria and eukaryotic algae some of the results from chapter 3 were included in the published study (Appendix 7.2). We found that the studied cyanobacteria have a preference for pseudo-cobalamin, while eukaryotic algae prefer cobalamin. In a bioinformatics study of more than 100 cyanobacteria, we showed that cyanobacteria are unlikely to have the capacity to produce cobalamin, but the majority of them can produce pseudo-cobalamin. Cyanobacteria are therefore unlikely to be the source of cobalamin for eukaryotic algae in the oceans.


The red alga Porphyra umbilicalis is part of an ancient group of algae and serves as a food source (laver). In the study I contributed to annotation of the 87.7 Mbp genome, sequenced by an international consortium of over 20 different research groups. Among the findings were that this algal genome contained both METE and METH suggesting that this organism can take up B12 from the environment when available but survive when it is not (Appendix 7.3).

The model system for algae bacterial interactions that was studied in chapter 4 was also the focal point of another study. Using quantitative proteomics, monocultures of *L. rostrata* supplemented with B12 and co-cultures of *L. rostrata* with *M. loti* were examined. From the study, 588 algal proteins were identified and quantified with 153 having different abundances when comparing the two cultures. We found that enzymes related to amino acid synthesis were expressed at higher levels in co-cultures while enzymes related to photosynthesis were expressed at lower levels. These observations were verified through physiological studies (Appendix 7.4).

2 Genome Scale
Metabolic Model of M. loti

2.1 Introduction

To study the microbial members of a productive and/or stable community, genomics approaches are being used, both to identify organisms and to characterise the specific interactions between the organisms. Metagenomics, that is sequencing the combined genomes of organisms in a mixed population (Handelsman et al, 1998), has become increasingly cheap and accessible as sequencing methods have developed in the past 15 years and has been used for identification of species (Sharon & Banfield, 2013). In one of the most cited metagenomics studies from the past decade, Qin et al, (2010) used metagenomics to identify and map the bacterial landscape of the human gut in 124 people. From the sequencing data they identified a set of bacterial species and genes that are thought to present in all humans, dubbed the minimal gut bacterial genome.

However, as the examples in section 1.2.2 illustrate, identifying the specific mechanisms and metabolic exchanges that underlie the interactions is still challenging and not fully uncovered for many consortia. Symbiotic interactions between microbial species can be underpinned by exchange of metabolites and using Genome Scale Metabolic (GSM) models has therefore been proposed as a way forward for illuminating these exchanges (Oberhardt et al, 2009; Zomorrodi & Maranas, 2012).

Metabolic capability is a central aspect when studying any microorganism. Through the emergence of next generation sequencing and exponentially decreasing cost the opportunity to create GSM models has become possible for
many organisms (Stein, 2010). GSM models are based on annotation of genes and characterisation of the enzymatic capabilities of the proteins encoded by the organism’s genes. They were first developed to study phenotypes of *Haemophilus influenzae* (Edwards & Palsson, 1999) and have since been used to study a variety of organisms as more have been genome-sequenced. Currently, over 160 GSM models for more than 100 species are available through online resources (Feist *et al*., 2009, Data updated in July/2017).

Table 2.1 - Examples of published GSM models. For each organism is summarized the number of genes accounted for by the model, the number metabolic reactions and the number of metabolites involved in those reactions.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Publication</th>
<th>Genes</th>
<th>Metabolic Reactions</th>
<th>Metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>(Edwards &amp; Palsson, 1999)</td>
<td>296</td>
<td>488</td>
<td>343</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>(Edwards &amp; Palsson, 2000)</td>
<td>660</td>
<td>627</td>
<td>438</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>(Förster <em>et al</em>, 2003)</td>
<td>708</td>
<td>1175</td>
<td>584</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>(Duarte <em>et al</em>, 2007)</td>
<td>1,496</td>
<td>3311</td>
<td>2233</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>(Poolman <em>et al</em>, 2009)</td>
<td>N/A</td>
<td>1406</td>
<td>1253</td>
</tr>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>(Chang <em>et al</em>, 2011)</td>
<td>1080</td>
<td>2190</td>
<td>1068</td>
</tr>
</tbody>
</table>

A GSM model for the plant *Arabidopsis thaliana* was first developed by Poolman *et al* (2009) and the model demonstrated that the principles and assumptions developed for bacteria models could be adopted and used to model eukaryotes. The *A. thaliana* model was later expanded by de Oliveira Dal'Molin *et al* (2010) to include compartment specific reactions and consider light uptake by the plant. In the study, it was demonstrated it is possible to simulate difference in redox reactions between a photosynthesis and non-photosynthetic state. Two GSM models for the green alga *C. reinhardtii* were published in 2011. Chang *et al* (2011) first published a model with a focus on lipid metabolism and modelling light uptake in the alga and used the model to develop a light source optimal for stimulating growth of the cell. A second model was published by Dal'Molin *et al* (2011) covering 866 genes and including 1862 metabolites and 1725 reactions based on the their *A. thaliana model*. This GSM model was focused on compartmentalization of reactions into the cytoplasm, mitochondrion, plastid and microbody.

GSM models have been applied for purposes such as guidance for metabolic engineering, and for studying interspecies interactions. Optimising productivity in
biological production systems is of key importance to the biotechnology industry. As random mutagenesis had not provided significant improvements for lycopene (a pigment) production, Alper et al (2005a) analysed the metabolic capabilities of *E. coli* using a GSM model and identified a set of gene knockout candidates that could increase the production in *E. coli*. The knockout strategy from the GSM model study was implemented by Alper et al (2005b), knocking out the genes *gdhA*, *aceE* and *fdhF*, and the resulting strain showed a 8.5-fold increase in lycopene production compared to the wild type. Another example of model guided metabolic engineering is the work by Izallalen et al (2008) to identify gene knockouts in *Geobacter sulfurreducens* that result in increased respiration for use in bioremediation.

There are multiple examples of GSM models being applied to study interspecies interactions. One of the first uses of GSM models to study mutualism was by Stolyar et al (2007) who investigated the interaction between *Desulfovibrio vulgaris* and *Methanococcus maripaludis*. In the symbiosis, *D. vulgaris* takes up lactate from the environment and secretes self-inhibiting bi-products that can be utilized as an energy source by *M. maripaludis*. The study focused on the interspecies electron transfer and showed that hydrogen was the main electron carrier between the organisms. The community model was created using individually developed models for each organism and the models were then combined as compartments inside a shared environment. (Zhang et al, 2011). An extension of this concept can help in optimising generation of a desired product. Ethanol production from sugars released from lignocellulose feedstocks is limited by sequential uptake of substrate in pure cultures of fermenting organisms, e.g. pentoses are taken up before hexoses. Hanly & Henson, (2011) therefore proposed to use a microbial community so that substrates could be used simultaneously. The authors used GSM models of *E. coli* and *S. cerevisiae* combined with kinetic data to investigate ethanol production from a combination of xylose and glucose. They applied the models to optimize inoculation ratio (*E. coli* to *S. cerevisiae*) and substrate ratio (glucose to xylose) and predicted a 2-fold increase in ethanol production compared with a pure culture of optimized *S. cerevisiae* grown on glucose. In a later study, they sought to validate the predictions experimentally and found improved ethanol yield of ~30% (Hanly et
al, 2012). Klitgord & Segrè, (2010) proposed a framework called Search for Exchange Metabolites that attempts to identify environments that induce interactions between organisms based on their GSM models. This approach was applied by Freilich et al (2011) to predict symbiotic and competing pairs of organisms among 118 bacterial species. In a study of 800 different microbial communities, GSM models of the community partners was used to demonstrate that co-occurrences of bacteria is likely to be a result of metabolic dependencies (Zelezniak et al, 2015).

2.1.1 GSM models and Flux Balance Analysis

GSM models are based on the genome of the target organism. Using annotations from the genome, enzymatic reactions are identified and the chemical reactions they catalyse are collected. The stoichiometric information related to each reaction is converted into a formal representation as one matrix with the dimensions $m \times n$, with $m$ being the number of metabolites and $n$ being the number of reactions. Each column in the table describes a specific reaction by the metabolites involved and their stoichiometry. Each row represents a specific metabolite and the reactions it is involved with. The stoichiometry of metabolite $i$ in reaction $j$ will be $n_{ij}$ and with $n_{ij} < 0$ for substrates and $n_{ij} > 0$ for products.

This matrix is referred to as a stoichiometric matrix ($S$) and the details on constructing the $S$ for an organism is elaborated in section 2.2.2. The changes in concentrations for metabolite $C$ can be described using the $S$ matrix as

$$\frac{dC}{dt} = S \times v(t) \quad (1)$$

$$v(t) = [v_1 \ldots v_n] \quad (2)$$

where $C$ is any metabolite, and $v(t)$ is a flux vector at a specific time point. Figure 2.1 shows an example of a simple metabolic model and the related S matrix. Metabolism operates on a relatively fast time scale compared to other cellular events such as translation, regulation or cell division. This is used to justify the central assumption that the cell is in pseudo steady state, where metabolite concentrations are constant and that all reactions occur at the same speed. This is formally written as
By making a steady state assumption it is possible to do a systems level analysis of metabolism without kinetic information for each reaction. While this is a simplification, it allows metabolic models to be expanded to a genome-scale level and changes the nature of the equations from a differential equation system to a linear equation system. The downside of the assumption is that metabolite concentrations are no longer variable, and the system is not directly able to simulate a dynamic situation.

$S \cdot \nu(t) = 0$  \hspace{1cm} (3)

Once created, GSM models can be analysed using Flux Balance Analysis (FBA). This applies linear programming to identify flux distributions across the model. For most GSM models, there are more reactions than metabolites. This causes
the system of equations to be undetermined (more variables than equations) and an infinite number of solutions to exist. It is therefore necessary to define a set of constraints on the model and provide an objective for the FBA. An example of a constraint is a fixed influx of glucose through a transporter reaction. This will limit the number of solutions to the model but can still generate multiple solutions. The model can therefore be provided with an objective, such as maximal export of a metabolite (e.g. B$_{12}$), maximal biomass production or minimal ATP production. When given constraints and objectives, the FBA framework will determine the optimal solution satisfying the conditions, with ‘optimal’ being defined by the objective function (example in Figure 2.2).

![Figure 2.2 – Example of flux distributions with constraints. The solutions were constrained by an influx of metabolite A of 1 unit, and the objective is maximal export of metabolite B. While both solutions are within the constraints, solution 1 gives a higher export of B and will therefore be identified as the best solution.](image)

### 2.1.2 Algae-bacteria model system

The aim of this study was to improve the understanding of mutualistic algae and bacteria relationships through studying the metabolism of the soil bacterium *M. loti*. *M. loti* is a member of a model system developed to study mutualistic alga/bacterium interactions. The other member of the system is the green alga *L. rostrata* (Kazamia *et al*, 2012b; Grant *et al*, 2014). The system is described in more detail in Chapter 1.

Kazamia *et al* (2012b) investigated the B12-dependence of *L. rostrata* by coculturing the alga with *M. loti* and *S. meliloti*, both of which can synthesise B12.
The study showed that when *L. rostrata* and *M. loti* were co-cultured in media without B12 and a fixed carbon source, both organisms grew, whereas this was not the case with *S. meliloti* unless a carbon source was provided. Table 2.2 provides a summary of the environments that support and do not support the growth of *L. rostrata* and *M. loti*.

Table 2.2 - Conditions that support symbiotic behaviour for *M. loti* and *L. rostrata*

In the left table is shown the conditions that support growth of *L. rostrata* and on the right is the conditions that support *M. loti* (Kazamia et al, 2012b; Grant et al, 2014)

<table>
<thead>
<tr>
<th>Condition</th>
<th>L. rostrata</th>
<th>M. loti</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axenic (minus B12)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Axenic (+ B12)</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>With <em>M. loti</em></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>With <em>S. meliloti</em></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Kazamia et al (2012b) found that the interaction between *M. loti* and *L. rostrata* has a stable equilibrium. In Figure 2.3A it is shown that with initial inoculation ratios varying from $10^{-6}$ to $10^5$ bacteria to algae, the co-culture converges towards a steady ratio of approximately 30 bacteria per alga. It should also be noted that this equilibrium is stable over time, and under different growth conditions. The B12 component of the symbiosis was further investigated by Grant et al (2014) who found that B12 per *M. loti* cell increases 10-fold when the bacterium is co-cultured with *L. rostrata* (Figure 2.3B).
2.1.3 **Aim of this study**

The aim of this study was to increase the understanding of *M. loti* metabolism and identifying metabolites or enzymatic reactions that allow it to form a stable mutualistic interaction with *L. rostrata*. I wanted to develop a GSM model representing the *M. loti* metabolism. Then, using a model of the *S. meliloti* metabolism, I carried out a comparison of the two organisms on a metabolism level to identify differences that could help to identify the metabolites exchanged between *S. meliloti* and *L. rostrata*. Using the developed GSM model for *M. loti*, I also wanted to investigate if the increased B12 production by the bacterium could be traced back to *L. rostrata* providing a particular precursor related to B12.

2.2 **Methods**

2.2.1 **Strains and genome sequences**

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>Genome sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. loti</em></td>
<td>MAFF303099</td>
<td>(Kaneko et al, 2000)</td>
</tr>
<tr>
<td><em>S. meliloti</em></td>
<td>1021</td>
<td>(Galibert et al, 2001)</td>
</tr>
</tbody>
</table>

2.2.2 **Creation of a genome scale metabolic reconstruction**

The process of creating a GSM model is described in a protocol by Thiele & Palsson, (2010) and forms the basis of my approach. The methods used were further developed through personal communication with Dr. Thomas Pfau,
University of Luxembourg. The process is a combination of automated and manual steps to take a genome and convert it into a set of metabolic reactions that represent the metabolism of the target organism.

### 2.2.2.1 Software packages and databases

Below is a brief description of the software packages and versions of these used for this project. The functionality of the software packages is described in detail when they are used.

Pathway tools version 13 (Karp et al, 2010) was used to assemble the initial draft genome. The manual curation steps were aided by the metabolism databases BioCyc (Caspi et al, 2012) and KEGG (Kanehisa et al, 2014). To aid the manual curation, several python scripts were written (Python Software Foundation, 2011). The details of these scripts are not described in this document as their functions and structures are basic. FBA was performed using the software suite ScrumPy version 731 by Poolman (2006). ScrumPy was selected as it was developed with the specific purpose of doing FBA and has a suitable method for integrating data from the BioCyc Database.

### 2.2.2.2 Assembly of draft model

The initial draft for the model of *M. loti* was created using an automated approach based on the annotated genome of *M. loti* MAFF303099 (Kaneko et al, 2000). The genome is available at [http://www.ncbi.nlm.nih.gov/genome/?term=txid266835](http://www.ncbi.nlm.nih.gov/genome/?term=txid266835). The automated approach for draft model creation used PathoLogic from the Pathway Tools package (Karp et al, 2010). The module takes the annotated genome of the target organism as input. Terms such as names, Enzyme Commission (EC) numbers and descriptions from the annotated genome are extracted and compared to the MetaCyc Database (Caspi et al, 2012). An example of this process is shown in Table 2.4. An inherent problem with this approach is that a missing or wrong annotation in the genome will lead to a missing reaction and therefore gaps in related pathway. To compensate for this, a workflow for filling these gaps was
applied. Based on the reactions and pathways added in the first step, gaps in the pathways were identified and using the sequence homology search tool BLAST, reactions to fill these gaps were identified and added to the model. Counter measures such as the taxonomy of the target organism and size of pathway were used to minimize the number of false positives reactions included in the model.

Table 2.4 – Example of the automated annotation process for draft assembly. In the example, the M. loti gene mlr1387 has been annotated as cobyrinic acid a,c-diamide synthase. Based on this annotation, the EC 6.3.5.9 is identified. The EC number is then correlated to MetaCyc ID R341-RXN. This reveals that mlr1387 is part of the B12 synthesis pathway and is responsible for the reaction shown.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Id</td>
<td>mlr1387</td>
</tr>
<tr>
<td>Gene annotation</td>
<td>cobyrinic acid a,c-diamide synthase</td>
</tr>
<tr>
<td>EC Number</td>
<td>6.3.5.9</td>
</tr>
<tr>
<td>MetaCyc ID</td>
<td>R341-RXN</td>
</tr>
<tr>
<td>Pathway</td>
<td>Adenosylcobalamin biosynthesis II (late cobalt incorporation)</td>
</tr>
<tr>
<td>Reaction</td>
<td>hydrogenobyrinate + 2 L-glutamine + 2 ATP + 2 H2O → 2 L-glutamate + hydrogenobyrinate a,c-diamide + 2 ADP + 2 phosphate + 2 H+</td>
</tr>
</tbody>
</table>

### 2.2.2.3 Reactions not suitable for GSM models

Using FBA to analyse the GSM model imposes certain restrictions on the reactions allowed in the model. Reactions are in general restricted to monomers, as all reactions must be specific. If polymers were allowed, new reactions would have to be added for each variant of the oligomer, e.g. ADP-D-Glucose + 1-4-alpha-D-Glucanₙ → ADP + 1-4-alpha-D-Glucanₙ₊₁ with \( n \) ranging from 1 to infinity. To reduce the number of reactions, reactions related to polymers are taken out of the draft model. A key feature of a GSM model is mass balance across the model, and certainty that all elements are accounted for. This leads to the restriction that all reactions need to be specific, and non-specific reactions are therefore removed from the draft model. The non-specific reactions in a draft model are a result of poor or weak annotation where general names are used as annotation. Reactions related to macromolecules, polymers and non-specific metabolites were identified during the stoichiometric balancing and were removed as part of the curation process.

### 2.2.2.4 Stoichiometric balancing

The reactions present in the model were inspected one by one and their stoichiometry was balanced on an elemental level manually. The chemical composition for metabolites was extracted from BioCyc. For metabolites without a chemical structure, symbolic structures were defined and used to balance
reactions. In the example shown in Figure 2.4, one of the metabolites, “Quinones” has no defined chemical structure. Instead R is used to represent the structure on both sides of the equation.

---

**Reaction:** Malate dehydrogenase

Phenylacetyl-CoA + 2 Quinones + H₂O → Phenylglyoxylyl-CoA + 2 Reduced-Quinones

**Stoichiometric definitions**

- Phenylacetyl-CoA - C₂₉H₃₈N₇O₁₇P₃S
- Quinones - R₁
- Phenylacetyl-CoA - C₂₉H₃₆N₇O₁₈P₃S
- Reduced-Quinones – R₁H₂

Figure 2.4 – Stoichiometric balancing of a reaction. In the example a metabolite does not have a chemical formula. In this case, the chemical formula of the Quinone is represented as R₁ and R₁H₂ in the reduced form.

---

### 2.2.2.5 Thermodynamic feasibility

To identify and remove thermodynamically unfeasible cycles, a set of FBA simulations was performed. A flux through the energy producing ATPase reaction (ATP + H₂O → ADP + Pi + H⁺) was set at 1 mmol / gDW / h. At the same time, the model was constrained to no external energy sources (e.g. glucose, sucrose, maltose etc.). With these conditions, any solution identified would mean that the model would be violating thermodynamics. Identified solutions provided a list of reactions that caused the model to violate thermodynamics. The list of reactions was inspected manually, and selected reactions were changed with respect to reversibility. This procedure was repeated whenever changes to the model were made, to ensure that no new unfeasible cycles were being introduced.

### 2.2.2.6 Biomass objective function

The biomass objective function (BOF) is an artificial reaction added to the model to simulate growth. The BOF is similar to other reactions in the model in the sense that it has substrates, products and a specific stoichiometry for each. The BOF simulates growth by consuming substrates, such as amino acids, nucleotides, ATP and co-factors to produce the product ‘biomass’ which is then exported from the model. The BOF used for the current GSM model of *M. loti* is an adapted version of the *E. coli* biomass composition described by Feist et al (2007). The biomass function used here was modified by removing a set of metabolites that are not relevant for the analysis performed in this work (Table 2.5).

The components in the biomass function served to guide the model development process and were used to measure progress. One by one, the biomass precursors required for growth were inspected and it was ensured that the model
contained reactions that could produce the appropriate set of metabolites from glycerol, NH₄, SO₄, P, O₂ and CO₂. If the model did not contain the necessary reactions to produce a specific biomass precursor, the pathways related to that compound were investigated manually using KEGG and MetaCyc (Caspi et al., 2012). The reactions required to complete all the pathways needed for biomass production were manually added to model. As B12 is of particular interest to the project, the B12 biosynthetic pathway was curated extensively and compared with genetic data to evaluate inclusion in the model. B12 is not part of the biomass function for E. coli and was therefore added to the M. loti biomass function. The B12 required per cell was estimated using the dry weight per cell for E. coli as shown in Table 2.6.
Table 2.5 - The Biomass Objective Function used for the *M. loti* and *S. meliloti* models. The function is used to simulate consumption of metabolites as part of growth. The number next to each metabolite shows the amount consumed per unit of biomass created, e.g. 0.5137 g of alanine is consumed per 1 g of biomass created.

<table>
<thead>
<tr>
<th>Nucleotides</th>
<th>Amino acids</th>
<th>Soluble and inorganic metabolites</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>dTTP</td>
<td>glycine</td>
<td>H$_2$O</td>
<td>ADP</td>
</tr>
<tr>
<td></td>
<td>L-tryptophan</td>
<td>pyridoxal 5'-phosphate</td>
<td>59.81</td>
</tr>
<tr>
<td>dGTP</td>
<td>L-cysteine</td>
<td>di-trans,octa-cis-undecaprenyl</td>
<td>H$^+$</td>
</tr>
<tr>
<td></td>
<td>L-histidine</td>
<td>diphosphate</td>
<td>57.129029</td>
</tr>
<tr>
<td>dATP</td>
<td>L-tyrosine</td>
<td>10-formyl-tetrahydrofolate</td>
<td>diphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>L-methionine</td>
<td>tetrahydrofolate</td>
<td>Inorganic</td>
</tr>
<tr>
<td>GTP</td>
<td>L-phenylalanine</td>
<td>5,10-methylenetetrahydrofolate</td>
<td>phosphor</td>
</tr>
<tr>
<td>CTP</td>
<td>L-serine</td>
<td>S-adenosyl-L-methionine</td>
<td>59.806</td>
</tr>
<tr>
<td>UTP</td>
<td>L-proline</td>
<td>coenzyme B12</td>
<td></td>
</tr>
<tr>
<td>ATP</td>
<td>L-asparagine</td>
<td>riboflavine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-aspartate</td>
<td>coenzyme A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-threonine</td>
<td>protoporphyrin IX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-glutamine</td>
<td>FAD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-glutamate</td>
<td>NADP$^+$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-isoleucine</td>
<td>NAD$^+$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-arginine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-lysine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-valine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-leucine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-alanine</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.6 – Estimation of B12 content per unit of biomass in *M. loti*

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>g B12 / <em>M. loti</em> cell</td>
<td>$3 \times 10^{-19}$</td>
<td>(Grant <em>et al.</em>, 2014)</td>
</tr>
<tr>
<td>E. coli cells / g Dw</td>
<td>$3.57 \times 10^{12}$</td>
<td>(Neidhardt &amp; Curtiss, 1996)</td>
</tr>
<tr>
<td>mmol B12 / g B12</td>
<td>0.7374</td>
<td></td>
</tr>
<tr>
<td>mmol B12 / g DW</td>
<td>$3 \times 10^{-19} \times$</td>
<td>$= 7.9 \times 10^{-7}$ mmol B12 / g DW</td>
</tr>
<tr>
<td></td>
<td>$3.57 \times 10^{12} \times$0.7374</td>
<td></td>
</tr>
</tbody>
</table>

### 2.2.3 *In silico* experiments

Using FBA and the ScrumPy modelling environment, a set of *in silico* experiments were carried out. The protocols used for these experiments are described below.

**2.2.3.1 Vitamin B12 and ATP yield from various substrates**

The methodology applied is based on work by Blank *et al* (2005). A list of all metabolites in the model was created and reactions for uptake of each of the metabolites were added. The import of metabolites was set to zero. This restriction was lifted for one reaction at a time and free import was allowed. At the same time the flux through the B12 exporter reaction was fixed to 1 mmol / gDw / h. The objective of the model was set to minimize the total intracellular flux. The same protocol was used to analyse ATP yields from substrates.

**2.2.3.2 Growth supporting substrates**

In this experiment both the *M. loti* model and a *S. meliloti* model developed by T Pfau (2013) were used to perform a comparative study between the two organisms. A list of metabolites present in both models was compiled and the metabolites ability to support growth in both organisms was tested. Importer reactions for the substrates were added to both models and the flux of all other import of organic metabolites was set to zero. Consecutively, the metabolites from the list were allowed as input, while a flux through the biomass function was set to one mmol / gDw / h. If a solution was found, the metabolite was designated as growth supporting. NH4, SO4, P, O2 and CO2 were freely available in the simulations.

**2.2.4 Experimental validation of growth supporting substrates**

Growth of *M. loti* and *S. meliloti* on different carbon sources was evaluated using the BIOLOG GN-2 kit (Garland & Mills, 1991). Single colonies were taken from TY plates
and suspended in an inoculation fluid provided as part of the kit. Aliquots of 100 µL were transferred from the inoculation fluid to each of the 96 wells in the BIOLOG microtiter plate, which has a range of different carbon sources in each well. The plate was incubated at 28 ºC for 7 days before growth was evaluated. Each well in the plate contained a rich medium that supports the incubated organisms as well as a dye bound to the carbon source that turns from transparent to purple if NADH is formed by the organism, indicating respiratory activity. A substrate was designated as growth supporting if any purple colour was observed in the assay. No quantitative evaluation of growth was performed.

2.3 Results
The model systems of *L. rostrata* and *M. loti* exhibit mutualistic behaviour when grown in co-culture (Kazamia et al., 2012b). It has been established that *M. loti* provides B12, which is essential for *L. rostrata* growth and that *L. rostrata* provides an organic energy source derived from photosynthate to *M. loti*. It has also been shown that *M. loti* increases its B12 production approximately 10-fold when grown in co-culture with *L. rostrata* compared to an axenic culture using glycerol as a carbon source (Grant et al., 2014), although this may simply be as a result of the algal cells acting as a sink. By developing a GSM model of *M. loti* and applying the analytical method of FBA to the model, I attempted to provide insights into the production of B12 under different conditions with the aim of illuminating the mechanisms and exchanges underpinning the symbiosis.

2.3.1 GSM model of *M. loti*
Based on the annotated genome of *M. loti*, PathwayTools/PathoLogic and the metabolism database BioCyc, an initial draft model was automatically generated as described in section 2.2.2.2. It contained 1908 reactions and 1804 metabolites. As the draft model was created using an automated approach, it was expected to contain several reactions not needed or suitable for the subsequent FBA. A set of manual curation steps was therefore applied to the draft to create the functional GSM model as described in section 2.2.2.3, 2.2.2.4 and 2.2.2.5.
2.3.1.1 Removal of irrelevant reactions

GSM models and FBA require that the metabolites of each reaction are specific, hence have defined substrates and products. General annotations of genes lead to the non-specific reactions being included in the first interaction of the model. In Table 2.7 are two examples of reactions that were removed as part of the curation process. A total of 537 reactions were removed in this step.

Table 2.7 – Examples of reactions from the initial draft that were removed as part of the curation process because they did not have defined substrates or products and therefore could not be used for the FBA. Reactions like these were identified through systematic inspection of each reaction in the initial model

<table>
<thead>
<tr>
<th>Gene</th>
<th>mll8195</th>
<th>Annotation</th>
<th>alcohol dehydrogenase</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC number</td>
<td>1.1.2.7</td>
<td>Reaction</td>
<td>a primary alcohol + NAD+ &lt;=&gt; an aldehyde + NADH + H+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>mll5109</th>
<th>Annotation</th>
<th>alpha-glucosidase</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC number</td>
<td>2.4.1.18</td>
<td>Reaction</td>
<td>(1,4-α-D-glucosyl) + H2O &lt;=&gt; (1,4-α-D-glucosyl) + α-D-glucose</td>
</tr>
</tbody>
</table>

2.3.1.2 Stoichiometric balancing and thermodynamic inconsistency

Stoichiometric balancing on an elemental level is a basic requirement for any metabolic model. Any violation of this will make subsequent predictions irrelevant. A thorough investigation was performed using the chemical composition of each metabolite. It was determined that there were no errors in the stoichiometry in any reaction in the model.

Thermodynamically unfeasible cycles are known to arise when creating a GSM model. These futile cycles carry a flux through a set of reactions with net conversion of metabolites. They typically arise from a lack of knowledge of the reversibility of many reactions and from the assumption that enzymatic reactions are reversible in a biological system unless there is any evidence to the contrary (Schellenberger et al, 2011). As part of the curation step, futile cycles that could replenish ATP without any uptake from the environment were identified and corrected. An example of a futile cycle and the subsequent correction is shown in Figure 2.5. The futile cycles were corrected using information about Gibbs free energy under standard conditions (pH 7, 1atm, 25°C), and reversibility was changed to irreversibility accordingly. The reversibility of eight reactions (Table 2.8) was changed to eliminate thermodynamic violating reactions in this model.
Figure 2.5 – Example of a cycle of reactions violating thermodynamics.
In A is shown the futile cycle and B shows the corrected form where no cycles exist. The reversibility of Formate:tetrahydrofolate ligase was changed as shown in Table 2.8 to remove the cycle.
<table>
<thead>
<tr>
<th>MetaCyc ID</th>
<th>MetaCyc Reaction</th>
<th>Gibbs Energy [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>RXN-0</td>
<td>&quot;PUTRESCINE&quot; + &quot;ACETYL-COA&quot; -&gt; &quot;PROTON&quot; + N-acetylpuroescine&quot; + &quot;COA&quot;</td>
<td>-745.4386</td>
</tr>
<tr>
<td>3.6.5.1-RXN</td>
<td>GTP + &quot;WATER&quot; -&gt; &quot;PROTON&quot; + &quot;GDP&quot; + &quot;PI&quot;</td>
<td>-102.58893</td>
</tr>
<tr>
<td>FORMATETHFLIG-RXN</td>
<td>ATP + &quot;FORMATE&quot; + &quot;THF&quot; -&gt; &quot;ADP&quot; + [&quot;PI&quot;] + &quot;10-FORMYL-THF&quot;</td>
<td>-25.441711</td>
</tr>
<tr>
<td>RXN-9929</td>
<td>DI-H-OROTATE + &quot;FUM&quot; -&gt; &quot;OROTATE&quot; + &quot;SUC&quot;</td>
<td>-20.54004</td>
</tr>
<tr>
<td>1.2.1.9-RXN</td>
<td>GAP + &quot;NADP&quot; + &quot;WATER&quot; -&gt; 2 &quot;PROTON&quot; + &quot;G3P&quot; + &quot;NADPH&quot;</td>
<td>-16.5318</td>
</tr>
<tr>
<td>RXN0-3962</td>
<td>ACETALD + &quot;NADP&quot; + &quot;WATER&quot; -&gt; &quot;ACET&quot; + &quot;NADPH&quot; + 2 &quot;PROTON&quot;</td>
<td>-16.52362</td>
</tr>
<tr>
<td>RXN-969</td>
<td>GLYCOLLATE + &quot;OXYGEN-MOLECULE&quot; -&gt; &quot;GLYOX&quot; + &quot;HYDROGEN-PEROXIDE&quot;</td>
<td>-21.019997</td>
</tr>
<tr>
<td>RXN-9623</td>
<td>PALMITATE + &quot;CO-A&quot; + &quot;ATP&quot; -&gt; &quot;PALMITYL-COA&quot; + &quot;PPI&quot; + &quot;AMP&quot;</td>
<td>-80.04077</td>
</tr>
<tr>
<td>RXN-8092</td>
<td>ACETALD + &quot;OXYGEN-MOLECULE&quot; + &quot;WATER&quot; -&gt; &quot;PROTON&quot; + &quot;ACET&quot; + &quot;HYDROGEN-PEROXIDE&quot;</td>
<td>-44.25712</td>
</tr>
</tbody>
</table>

### 2.3.1.3 B12 and biomass precursors pathway curation

B12 production in *M. loti* plays a central role among the mechanisms underpinning mutualism in the co-culture with *L. rostrata*. The B12 synthesis pathway therefore underwent thorough curation. See section 3.1.1 for a thorough review of B12 synthesis and the related genes. When investigating and curating the B12 synthesis pathway, it was discovered that both early and late cobalt insertion pathway had been added in the initial draft model due to sequence similarity between the respective genes. By analysing the genome sequence and annotations of *M. loti*, combined with pathway information from Warren *et al* (2002), the reactions without a genetic basis were removed and the pathways corrected accordingly. This left only the late cobalt insertion pathway functional.

The biomass function is used to simulate growth and consumption of a defined set of metabolites that serves as building blocks for the target organism. In order to make predictions related to growth, such as gene lethality, there should be a complete pathway so that each of the biomass precursors can be synthesised from basic nutrients such as glycerol, NH\(_4\), SO\(_4\), P, O\(_2\) and CO\(_2\). One by one the biomass precursors were investigated, and reactions were added if needed. Seven reactions that had been taken out in a previous curation step were added back in a corrected
form, and eight new reactions were added to the model, to allow for production of biomass Table 2.9.

<table>
<thead>
<tr>
<th>MetaCyc ID</th>
<th>Reaction</th>
<th>Gene in M. loti</th>
</tr>
</thead>
<tbody>
<tr>
<td>RXN-8627</td>
<td>(R)-1-aminopropan-2-ol + ATP -&gt; (R)-1-amino-2-propanol O-2-phosphate + ADP</td>
<td>mll9228</td>
</tr>
<tr>
<td>RXN-8626</td>
<td>L-threonine + ATP -&gt; L-threonine 3-O-phosphate + ADP + H+</td>
<td></td>
</tr>
<tr>
<td>GLYCOLALD-DEHYDROG-RXN</td>
<td>glycolaldehyde + NAD+ + H2O -&gt; glycolate + NADH + 2 H+</td>
<td>mll1168, mll6098</td>
</tr>
<tr>
<td>325-BISPHOSPHATE-NUCLEOTIDASE-RXN</td>
<td>adenosine 3',5'-bisphosphate + H2O -&gt; AMP + phosphate</td>
<td>mll7577</td>
</tr>
<tr>
<td>4.1.1.81-RXN</td>
<td>L-threonine 3-O-phosphate + H+ -&gt; (R)-1-amino-2-propanol O-2-phosphate + CO2</td>
<td>mll5890</td>
</tr>
<tr>
<td>AMINOPROPDEHYDROG-RXN</td>
<td>(R)-1-aminopropan-2-ol + NAD+ - aminoacetone + NADH + H+</td>
<td>mll3070</td>
</tr>
<tr>
<td>TRIPHOSPHATASE-RXN</td>
<td>PPi + H2O -&gt; phosphate + diphosphate</td>
<td>mll9114</td>
</tr>
<tr>
<td>DARAB5PISOM-RXN</td>
<td>D-arabinose 5-phosphate &lt;-&gt; D-ribulose 5-phosphate</td>
<td>mll3023</td>
</tr>
</tbody>
</table>

### 2.3.1.4 Curated GSM model of M. loti

The curated GSM model of *M. loti* produced in this work accounts for 1011 ORFs and represents the activities that these genes carry out. The model contains 1368 reactions and 1293 metabolites, which is comparable to the *S. meliloti* model developed by Pfau (2013) as indicated in Table 2.10. Both bacteria have considerably larger gene contents than *E. coli*, but even so, their GSM models include a similar number of genes and reactions as the model for *E. coli*. The model of *M. loti* can be found on [https://github.com/ulrichkudahl/PhDthesis/tree/master/GSM_Mesorhizobium_Loti](https://github.com/ulrichkudahl/PhDthesis/tree/master/GSM_Mesorhizobium_Loti).
Table 2.10 – Comparison of the GSM model for *M. loti* with *E. coli* and *S. meliloti*. Total genes refer to the number of genes in the genome. Included genes is the number of genes use to get the metabolic reactions. Metabolites and reactions refer to numbers from the model of the organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Total genes</th>
<th>Included genes</th>
<th>Metabolites</th>
<th>Reactions</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. loti</em></td>
<td>7,334</td>
<td>1011</td>
<td>1293</td>
<td>1369</td>
<td>This work</td>
</tr>
<tr>
<td><em>S. meliloti</em></td>
<td>6,363</td>
<td>Unknown</td>
<td>1215</td>
<td>1318</td>
<td>(Pfau, 2013)</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4,516</td>
<td>1260</td>
<td>1039</td>
<td>2077</td>
<td>(Feist <em>et al.</em>, 2007)</td>
</tr>
</tbody>
</table>

2.3.2 Assessment of growth supporting substrates

Kazamia *et al* (2012b) initially found that *L. rostrata* can support the growth of *M. loti* effectively, but not *S. meliloti*. Identifying carbon sources that support growth of *M. loti*, but not *S. meliloti* could provide candidate substrates for further experiments that might explain this difference. Using this approach 1018 metabolites present in both the *M. loti* and *S. meliloti* model were tested for their ability to support biomass production. A total of 168 substrates supported growth of *M. loti* and 248 supported growth of *S. meliloti*. Sixteen substrates were found to support growth of *M. loti*, but not *S. meliloti* (Table 2.11). Four of the substrates were tested experimentally using BIOLOG plates, as described in section 2.2.4. Despite predictions showing that the substrates only support growth of *M. loti*, the experimental data points to either growth or no growth of both organisms for each substrate.
Table 2.11 – Comparison of *in silico* and *in vivo* growth supporting substrates. Metabolites predicted to support experimental growth of *M. loti*, but not *S. melliloti* were tested in a BIOLOG growth assays. + indicates growth support and – indicates not supporting growth. For the four substrates tested so far and no differences were found between the two organisms.

<table>
<thead>
<tr>
<th>Metabolite predicted to support <em>M. loti</em> but not <em>S. melliloti</em></th>
<th>Experimental growth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M. loti</em></td>
</tr>
<tr>
<td>N-formimino-L-glutamate</td>
<td>+</td>
</tr>
<tr>
<td>Acrylamide</td>
<td></td>
</tr>
<tr>
<td>Acrylate</td>
<td></td>
</tr>
<tr>
<td>N-acetylputrescine</td>
<td></td>
</tr>
<tr>
<td>Aminoacetonate</td>
<td></td>
</tr>
<tr>
<td>4-acetamidobutanoate</td>
<td></td>
</tr>
<tr>
<td>(S)-lactate</td>
<td>-</td>
</tr>
<tr>
<td>4-acetamidobutanal</td>
<td></td>
</tr>
<tr>
<td>Putrescine</td>
<td>-</td>
</tr>
<tr>
<td>orotidine-5'-phosphate</td>
<td></td>
</tr>
<tr>
<td>Sucrose</td>
<td>+</td>
</tr>
<tr>
<td>acrylonitrile</td>
<td></td>
</tr>
<tr>
<td>2-(α-hydroxyethyl)thiamine diphosphate</td>
<td></td>
</tr>
<tr>
<td>L-asparagine</td>
<td>-</td>
</tr>
<tr>
<td>L-lactaldehyde</td>
<td></td>
</tr>
<tr>
<td>uridine-5'-monophosphate</td>
<td></td>
</tr>
</tbody>
</table>

2.3.3 B12 yield from various carbon sources in *M. loti*

Grant *et al* (2014) reported that *M. loti* produced 10-fold more B12 per cell in co-cultures compared with axenic cultures grown with glycerol as carbon source. Since the carbon source provided by *L. rostrata* to *M. loti* is currently not known, identifying substrates that have a higher maximal B12 yield than glycerol could help to identify the carbon source provided by *L. rostrata*. The hypothesis behind the experiment is that if a substrate is few conversion steps from the B12 precursor, such as uroporphyrinogen-III, it will have a higher B12 yield than a substrate that is many conversions away from a B12 precursor. The 1368 metabolites included in the *M. loti* were therefore tested *in silico* for their ability to support B12 production and the maximal B12 yield for each was calculated. Of the tested substrates 212 could support B12 production and the five substrates with the highest and lowest calculated yields are shown in Figure 2.6A. There is approximately a 10-fold difference between the substrates with the lowest and highest B12 / substrate yield, and a 5-fold difference between glycerol and the substrates with the highest yield.
To account for differences in ‘energy’ density between the substrates the maximal ATP yield for each substrate was also calculated. The ATP yield and B12 yields were correlated and plotted in Figure 2.6B. The figure shows that there is some level of correlation with a few substrates as outliers. All outliers except for one were substrates that had a proportionally higher ATP yield than B12 yield and are thus not of interest. The substrate 1-pyrroline-4-hydroxy-2-carboxylate stands out as having the highest B12 yield / ATP yield.
Figure 2.6 – Analysis of B12 yield from various substrates
A) The five substrates with the highest and lowest B12 yield and Glycerol which was used as substrate in the experiments by Grant et al (2014).
B) Correlation between B12 yield and ATP yield from 212 substrates that supported B12 production. ATP yield is shown in the x-axis, with B12 yield on the Y-axis. Interesting substrates would be those that have high B12 yield that cannot be explained by a high ATP yield.
2.4 Discussion

The symbiosis between *M. loti* and *L. rostrata* is a model system for symbiosis between a bacterium and an alga. *L. rostrata* is dependent on B12 as a co-factor for methionine synthase and is unable to grow without it. The B12 producing bacterium *M. loti* can produce and supply B12 to *L. rostrata*, which in turn provides the bacterium with a fixed carbon source derived from photosynthesis (Croft *et al*, 2005; Kazamia *et al*, 2012). The model system has been found to be stable over time and reaches an equilibrium of ~30 bacteria to one alga (Kazamia *et al*, 2012). It has also been shown that there is a ten-fold increase in B12 per *M. loti* cell when the bacterium is grown in co-culture with *L. rostrata*, compared with an axenic culture of the bacterium (Grant *et al*, 2014). While it is known that the bacteria produce and provide B12 to the algae, the metabolites that the algae contribute to the symbiosis are currently not known.

In this study, I attempted to use modelling of *M. loti*’s metabolism to identify metabolites that *L. rostrata* might supply to *M. loti*. Initially the goal of the study was to create models of both organisms and join them together to allow simulation of the community. As the project moved along, it became apparent that GSM models were not a suitable for simulating exchange co-factors. Co-factors such as B12 are not being consumed like other metabolites, as they are catalysts in reactions such as methionine synthase. There is therefore no stoichiometric change in the co-factor molecule and thus it is not suitable to include in GSM models. The focus of the study was therefore changed to the model of *M. loti* and its metabolism.

A set of reactions were assembled based on the annotated genome of *M. loti* and served as the initial draft of the model. These reactions underwent a set of manual and semi-automated curation steps to make the model useful for FBA. A total of 537 reactions were removed as they were not suitable for FBA. Similarly, thermodynamics violating cycles were identified and removed, and pathways for biomass production were curated and completed. The curated model of *M. loti* accounts for 1011 genes, 1369 reactions and contains 1293 metabolites. It should be noted that among the reactions that are not associated with a gene are 111 transporter reactions that are needed to explain the model, and indeed the physiological activity of the bacterium, such as the ability to take up phosphate. Some may be simply diffusion, e.g. O2, CO2, whilst others are gene products that are currently unknown. The remaining reactions
without a directly associated gene are added to the model because they complete a pathway. No direct association with a gene should not be regarded as false reaction, but rather one that needs more thorough curation. For some of the reactions it will be the case that they should be associated with a gene that is wrongly annotated, annotated vaguely or not annotated. However, there is also a possibility that some of these reactions are not present in *M. loti*. As part of the model refinement, further investigations of these reactions should be carried out to determine whether these reactions should be included or not. The proton gradient across the cell membrane is crucial to synthesis of ATP and is important for all quantitative predictions. A systematic analysis and correction of charge balance for all reactions should therefore be carried out as part of the model refinement. Charge balancing could be done using a workflow similar to the stoichiometric balancing workflow. The charge of all metabolites would be defined and, following that, each reaction can be inspected one by one, and charge conservation can be evaluated and corrected as needed.

FBA was used to generate a list of candidate substrates that all were predicted to support growth of *M. loti* and not *S. meliloti*. This study was undertaken based on observations that *S. meliloti* could not be supported by *L. rostrata* (Kazamia *et al.*, 2012b). Four of the 16 substrates on the list have been tested experimentally, but the results did not align with the predictions. Initially, the aim was to continue with the experimental verification process, but while this was ongoing, new experiments performed by PhD student Frederick Bunbury, Department of Plant Sciences, University of Cambridge, showed that *L. rostrata* could support *S. meliloti*. The initial premise for the experiment was therefore changed and experiments were discontinued. For the four substrates where model prediction and experimental data were not in agreement, the observations should be used to guide further model refinement. By analysing the pathways related to metabolism of these substrates, corrections can be made to ensure that model and experimental data are in accordance. Making these corrections will enable better predictions in future *in silico* experiments.

Currently, the predictions related to growth of *M. loti* in the GSM model are performed using a biomass function for *E. coli*. Since the predictions related to growth in this project are not quantitative, this is not an issue as of now. However, if the aim was to
predict quantitative growth of *M. loti*, it would be necessary to determine the biomass composition for *M. loti*. While there is no single protocol for this, there are established protocols for the individual parts of the biomass, e.g. determining protein content, total DNA, etc (Feist & Palsson, 2010).

Currently, there is no confirmed explanation for the 10-fold increase in B12 per cell observed when *M. loti* is co-cultured with *L. rostrata*, compared to growth on glycerol. It was therefore hypothesized that the substrate provided by *L. rostrata* might be able to be converted to B12 at a higher mol/mol yield than glycerol. A total of 1368 metabolites were therefore analysed and their maximal B12 yield was calculated and the yields were normalised using the ATP yield for each substrate. The simulations did not provide a candidate substrate that stood out with regards to ‘B12 yield’ versus ‘ATP yield’. The preliminary conclusion is therefore that the pathway of a substrate to B12 is not directly responsible for the 10-fold increase in B12 per *M. loti* cell. It is however still a possibility that the increased B12 production is related to the substrate provided by *L. rostrata*, though the mechanism might be regulatory instead of pathway-related. Cheng et al (2014) showed that adding 5 mg/L of the plant derived small molecule rotenone (C23H22O6) to a culture of the B12-producing *Pseudomonas denitrificans* caused an increase from 48.28 (±0.62) mg/L to 54.70 (± 0.45) mg/L of B12. When rotenone was added, increased levels of phosphofructokinase and pyruvate kinase activity was seen, and an increase in the glycolytic pathway was proposed as the cause of the increased B12 levels.

To develop the GSM mode-based approach further, one possible extension would be to integrate transcriptomics data for *M. loti* grown in axenic and in co-culture. By integrating expression data with GSM models, it will be possible to create a model specific to the environment it is simulating, hence one model of *M. loti* for the axenic state, and another for the co-culture state. This could make predictions more accurate as it will be possible to define which reactions are active and inactive in each state. The methods for combining a GSM model and omics data are well established and protocols have been published (Blazier & Papin, 2012; Saha et al, 2014).

There is no genome model for *L. rostrata* publicly available. My initial idea was therefore to develop a GSM model of *L. rostrata* from transcriptomics data. This has been done in other studies and protocols for it have been published (Robertson et al,
2010; Grabherr et al, 2011). I would apply the methods that I used to create the GSM model for *M. loti* to the *L. rostrata* transcriptome to refine the model from an initial draft. An additional step of evaluating and defining compartmentalization of the reactions will be added to the process. This step is likely to draw on data from the already developed models of *C. reinhardtii* (Dal'Molin et al, 2011; Chang et al, 2011).

Once developed, the GSM model of *L. rostrata* could be combined with the model of *M. loti* to simulate mutualism. Using FBA reactions, and thereby genes, that are central to interaction between the two organisms could be studied; then through a systematic and exhaustive simulation of *in silico* gene knockouts, a list of genes essential to the symbiosis could be identified. This would then serve to further the understanding of the underlying mechanisms of symbiosis. However, for the reasons discussed above, the GSM model of *M. loti* and FBA is not ideal for simulating exchanges of co-factors and unlikely to provide insights about the mechanisms underpinning the *L. rostrata/M. loti* mutualism. The model can however be used for other studies related to general metabolism of the bacterium. Using the model, it will be possible to explore pathways and assess metabolic consequences of gene-knockouts on larger scale and at a pace otherwise not possible using only experimental data.
3 Survey of Vitamin B12 synthesis, usage and dependence in bacteria

Preface

It should be noted that part of the results in this chapter has already been published. The results related to B12 synthesis as well as DMB synthesis in cyanobacteria are published in Helliwell et al. (2016). The computational framework and method described in this chapter was also published in the study.

The results related to niacin and biotin auxotrophy in Roseobacter and Sulfitobacter will be published as part of a study by Cooper et al. (manuscript in preparation)

Data and the computational framework can be found on https://github.com/ulrichkudahl/PhDthesis/tree/master/B12_Survey.

3.1 Introduction

Vitamin B12 is one of the most complex metabolites in nature. In biological systems, it is an enzymatic co-factor for essential reactions such as methionine synthase and ribonucleotide reductase. The metabolic relevance of B12 and the related B12-dependent enzymes are described in more detail in Chapter 1.

3.1.1 Vitamin B12 synthesis

After Hodgkin et al., (1955) discovered the complex structure of B12 attention turned to how this complex metabolite is synthesised in nature. Using the model organisms Pseudomonas denitrificans and Salmonella enterica genes were isolated and cloned into plasmids and expressed in E. coli individually and their function confirmed using enzymatic assays (Roth et al., 1996). These studies revealed that there are two synthesis pathways for this vitamin each containing 21 enzymes converting
Uroporphyrinogen III into the bioactive cobalamin. The synthesis pathways are extensively reviewed by Warren et al., (2002) and outlined in Figure 3.1. The two pathways are called the aerobic and anaerobic pathway, referring to the fact that the former requires molecular oxygen for the ring contraction step. They are also referred to as the late cobalt insertion-pathway (aerobic) and the early cobalt insertion pathway (anaerobic) – here referring to the fact that in the latter the majority of intermediates already have the Co\(^{2+}\) ion inserted in the tetrapyrrole ring. Proteins from the aerobic pathway were initially discovered in \(P.\) denitrificans and given the prefix \(Cob\), while the proteins from the anaerobic pathway isolated in \(S.\) enterica were given the prefix \(Cbi\). However, this is complicated by the genes encoding the enzymes from the last steps of the anaerobic pathway using the prefix \(Cob\) rather than \(Cbi\) (see Figure 3.1). The result of this is a confusing naming structure, where two distinct proteins are both named \(\text{CobT}\). To reduce any ambiguities, any enzyme name that is used twice is given a subscript to enable distinction e.g. \(\text{CobT}_{\text{aerobic}}\) and \(\text{CobT}_{\text{anaerobic}}\). Uroporphyrinogen III is derived from glutamate and also serves as a precursor to heme, siroheme and F430 (Roth et al., 1996), and is converted to precorrin-2 by either CysG or CobA. At this point the synthesis pathways split and the subsequent corrin ring formation and cobalt insertion occurs in opposite order facilitated by different enzymes. Cob(II)urinate a,c-diamide is formed by CIA and the Costs complex, the remainder of the synthesis is done in similar manner, but still using different enzymes. In the final steps of the synthesis pathway, the nucleotide loop is assembled, and the lower axial ligand is attached.

Some species, such as \(E.\) coli, do not have the required genes to synthesis B12 \textit{de novo}, but still have B12-dependent enzymes. Bacteria such as this cover their B12 need by taking up the vitamin from the environment using the btuCDF transport system (Locher et al., 2002; Borths et al., 2005). In some species, it has been demonstrated that the btuCDF system can also facilitate uptake of other corrinoids such as cobinamide. \textit{Thermotoga lettingae} is not able to produce B12 \textit{de novo}, but has been shown to produce B12 when it is supplied with cobinamide (Butzin et al., 2013).
Figure 3.1 – Schematic overview of the two B12 synthesis pathways. Enzymes from the early cobalt insertion-pathway are shown in green and the late cobalt insertion-pathway are shown in red. In the early cobalt insertion pathway, cobalt is inserted by either CbiK, CbiX or CysG, while in the late insertion pathway, this is done by Costs.
3.1.2 B12 variants – cobalamin and pseudocobalamin

Different forms of B12 can be found in nature, including variants in the lower ligand, with the two of the most studied forms being cobalamin and pseudocobalamin (Figure 3.2). In the cobalamin form, the molecule has a lower ligand called 5,6-dimethylbenzimidazole (DMB). DMB can be synthesised from flavin through a reaction catalysed by the enzyme 5,6-dimethylbenzimidazole synthase, which is encoded by the bluB gene (Campbell et al., 2006; Taga et al., 2007). Recently an alternative anaerobic synthesis pathway for DMB synthesis was discovered in Eubacterium limosum from the Clostridiales order (Hazra et al., 2015). This operon has been named bzaABCDE. Cobalamin is the form of B12 that is bioactive in humans, while pseudocobalamin have been shown to be 500x less bioavailable to humans. This is caused by a weaker binding to the intrinsic factor, which facilitates absorption of B12 from the human gut (Stupperich & Nexø, 1991). Pseudocobalamin received little attention due the limited applications for human health, but interest has since then grown. In place of the DMB molecule, pseudocobalamin has an adenine molecule, which is thought to be universally available in bacteria (Taga & Walker, 2008).

![Figure 3.2 – Structures of cobalamin and pseudocobalamin. The lower ligand is highlighted in the coloured box. For cobalamin this lower ligand is DMB, while in pseudocobalamin, the ligand is replaced by adenine.](image)

3.1.3 Using comparative genomics to study B12 synthesis

B12 synthesis has been studied in individual species and currently experimental evidence for B12 synthesis is available in 40+ bacterial species (Table 3.8). That is however only a small fraction of the total number of bacterial species that are thought
to produce B12. At this point it was unknown how common or rare the B12-producer phenotype is and in which phylogenetic groups it is dominant. Due to logistical challenge in trying to gather thousands of species of bacteria and test these in a biological assay, and due to the advances in sequencing techniques, comparative genomics has been proposed as a way forward to answer these questions. Comparative genomics have twice previously been applied with the aim of studying B12 synthesis across all bacteria, but in both cases the studies have included fewer than 500 species of bacteria (Rodionov et al, 2003; Zhang et al, 2009). In more recent studies efforts have focused on a group of bacteria from a particular environment such as the human gut (Degnan et al, 2014a) or the oceans (Sañudo-Wilhelmy et al, 2014). Common for all but the study by Degnan et al, (2014a) is that these studies have used a subset of the genes related to B12 synthesis rather than all genes from both pathways.

3.1.4 Aim of the study
The aim of this study was initially to identify bacterial species capable of B12 synthesis. This was motivated by the lack of data on B12 synthesis in bacteria and for related projects it would be useful to have a resource for quickly assessing the B12 synthesis capabilities of a given bacterium. During the project, it was expanded as several new research questions arose.

- It is generally accepted that mammals are B12-dependent and that plants are B12-independent (Smith et al, 2007; Froese & Gravel, 2010). However, examples of bacteria with and without B12 auxotrophy have been identified (Sudarsan et al, 2006; Degnan et al, 2014b). It is currently unknown which of these two cases is the most common, and what the ratio between them is.
- As described above, some bacterial species have been found to make cobalamin and others are synthesise pseudocobalamin. I wanted to discover which of these two metabolites are most common in nature
- Given the preponderance of B12-dependent algae in marine environments, I wanted to investigate if B12 synthesis was more common among marine bacteria compared to bacteria in other environments
3.2 Methods

3.2.1 Representative sequences for relevant enzymes

A comparative genomics approach centred on homology search was selected as the method to investigate the research questions. To carry forward with this method it was required that representative and experimentally verified amino acid sequences were identified for all the relevant enzymes related to B12 synthesis and relevant enzymes utilizing B12 as a cofactor. By searching through published literature, representative gene sequences for each relevant enzyme were found and a reference to the experimental work verifying the sequences’ enzymatic activity is shown in Table 3.1 - Table 3.5. In addition to the NCBI ID and literature reference, I have added information about the chemical reaction facilitated by the related enzyme. As conserved domain analysis forms part of the search strategy, I identified conserved domains in the representative sequences using CD-search (Marchler-Bauer et al, 2015). This collection of experimentally verified sequences, complete with NCBI GI, serve as a resource not only for this study, but also for future studies of the B12 synthesis pathway or any of the other pathways described.

In addition to cobalamin biosynthesis, I extended the analysis to look for the prevalence of biosynthesis of two other vitamins, niacin (vitamin B3) and biotin (vitamin B7). The representative sequences for each enzyme are found in Table 3.6 and Table 3.7.

3.2.2 Genomes, Homology search, protein domain analysis, phylogenetic trees, environmental data and computational frameworks

Amino acid sequences from NCBI NR protein database (NCBI Resource Coordinators, 2016) was extracted (March 2016) and a local version of the database was created. Data from NCBI Taxonomy (NCBI Resource Coordinators, 2016) was added to allow for homology search for individual species. Using information from NCBI Taxonomy, it was possible to structure the database to collapse species that have been sequenced multiple times into a single representative translated genome. In order to exclude partially sequenced genomes, only species with a minimum of 2000 protein sequences translated from their genome were included in the study. In total 8424 eubacterial species were represented in the final dataset. Although many
species of *archaeobacteria* are known to produce B12, their biosynthetic pathways differ and so would complicate the analysis. BLASTP (Camacho *et al*, 2009) with a threshold of e-value < $10^{-10}$ was used to identify putative hits in each species. The closest match to the query sequence was subject to a domain analysis using CDD (Marchler-Bauer *et al*, 2015) searching the databases Pfam (Finn *et al*, 2014), the COGs collection (Huerta-Cepas *et al*, 2016), TIGRFAMs (Haft *et al*, 2013) and the NCBI Protein Clusters collection (NCBI Resource Coordinators, 2016). I compared the identified domains in each of the putative homologues with the expected domains for each query sequence and used a threshold of at least one domain match the list of expected domains.

As part of the results section for this chapter, a number of phylogenetic trees were generated. The trees were created using NCBI Taxonomy IDs, PhyloT (http://phylot.biobyte.de/) and iTOL (Letunic & Bork, 2016).

Data on the natural habitat of bacteria were extracted from the GOLD database (Reddy *et al*, 2015) and correlated with data from this study using NCBI taxonomy IDs. Environmental data was not available for all species.

Due to the number of enzymes and translated genomes involved in this study, I built an automated computational pipeline using custom python scripts to organise the flow of data from one analysis to the next. This also allowed me to run the analysis on multiple CPUs, which decreased the computational time by more than an order of magnitude. The source code for the software in the project can be found in the Appendix 7.1 and https://github.com/ulrichkudahl/PhDthesis/tree/master/B12_Survey.
<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Exp_Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>cobA_aerobe</td>
<td>116832</td>
<td></td>
<td>cd11642,TIGR01469,PRK06136,PLN02625</td>
<td>S-adenosyl-L-methionine + precorrin-1 ↔ S-adenosyl-L-homocysteine + precorrin-2</td>
</tr>
<tr>
<td>cobI</td>
<td>116840</td>
<td></td>
<td>TIGR01467,cd11645,PRK05990</td>
<td>precorrin-2 ↔ S-adenosyl-L-methionine ↔ S-adenosyl-L-homocysteine + precorrin-3A + H+</td>
</tr>
<tr>
<td>cobG</td>
<td>1352123</td>
<td></td>
<td>TIGR02435,pfam03460</td>
<td>precorrin-3A + NADH + oxygen + H+ ↔ precorrin-3B + NADPH + H2O</td>
</tr>
<tr>
<td>cobJ</td>
<td>116841</td>
<td></td>
<td>TIGR01466,pfam00590,cd11646,PRK05991,COG1010</td>
<td>precorrin-3B + S-adenosyl-L-methionine ↔ precorrin-4 + S-adenosyl-L-homocysteine + 2 H+</td>
</tr>
<tr>
<td>cobM</td>
<td>116844</td>
<td></td>
<td>TIGR01465,pfam00590,cd11641,COG2875</td>
<td>precorrin-4 ↔ S-adenosyl-L-methionine ↔ precorrin-5 + S-adenosyl-L-homocysteine + H+</td>
</tr>
<tr>
<td>cobF</td>
<td>116837</td>
<td></td>
<td>COG2243,TIGR02434,cd11643</td>
<td>precorrin-5 ↔ S-adenosyl-L-methionine + H2O ↔ precorrin-6A + S-adenosyl-L-homocysteine + acetate + 2 H+</td>
</tr>
<tr>
<td>cobK</td>
<td>54040789</td>
<td></td>
<td>COG2099,pfam02571,TIGR00715,PRK08057</td>
<td>precorrin-6B + NADPH ↔ precorrin-6A + NADPH + H+</td>
</tr>
<tr>
<td>cobL</td>
<td>116843</td>
<td></td>
<td>TIGR02467,TIGR02469,COG2242,cd11644</td>
<td>precorrin-7 ↔ S-adenosyl-L-methionine ↔ precorrin-8x + S-adenosyl-L-homocysteine + 2 H+</td>
</tr>
<tr>
<td>cobH</td>
<td>116839</td>
<td></td>
<td>COG2082,PRK08285,pfam02570,PRK05782</td>
<td>precorrin-8x + H+ ↔ hydrogenobyrinate</td>
</tr>
<tr>
<td>cob</td>
<td>116833</td>
<td></td>
<td>TIGR00379,pfam01656,COG1797</td>
<td>hydrogenobyrinate + 2 L-glutamine + 2 ATP + 2 H2O ↔ 2 L-glutamate + hydrogenobyrinate a,c-diamide + 2 ADP + 2 phosphate + 2 H+</td>
</tr>
<tr>
<td>cobN</td>
<td>231829</td>
<td></td>
<td>TIGR02257,pfam02514,cd10150,PRK05989,COG1429</td>
<td>hydrogenobyrinate a,c-diamide + Co2+ + ATP + H2O ↔ cob(I)urinate a,c-diamide + ADP + phosphate + 3 H+</td>
</tr>
<tr>
<td>cobS_aerobe</td>
<td>231833</td>
<td></td>
<td>TIGR01650,pfam02654</td>
<td>hydrogenobyrinate a,c-diamide + Co2+ + ATP + H2O ↔ cob(I)urinate a,c-diamide + ADP + phosphate + 3 H+</td>
</tr>
<tr>
<td>cobT_aerobe</td>
<td>231834</td>
<td></td>
<td>TIGR01651,pfam06213,COG4547,pfam06213,pfam11775</td>
<td>hydrogenobyrinate a,c-diamide + Co2+ + ATP + H2O ↔ cob(I)urinate a,c-diamide + ADP + phosphate + 3 H+</td>
</tr>
<tr>
<td>cobR</td>
<td>17982642</td>
<td></td>
<td>No conserved domains</td>
<td>2 cob(I)urinate a,c-diamide + FMN + 3 H+ ↔ 2 cob(I)urinate a,c-diamide + FMNH2</td>
</tr>
<tr>
<td>cobO</td>
<td>231830</td>
<td></td>
<td>TIGR00708,pfam02572,cd100561,PRK05986</td>
<td>2 cob(I)urinate a,c-diamide + FMN + 3 H+ ↔ 2 cob(I)urinate a,c-diamide + FMNH2</td>
</tr>
<tr>
<td>cobQ</td>
<td>231832</td>
<td></td>
<td>TIGR00313,pfam01656,pfam07685,COG1492</td>
<td>adenosyl-cobinamide a,c-diamide + 4 L-glutamine + 4 ATP + 4 H2O ↔ 4 L-glutamate + adenosylcobinate + 4 ADP + 4 phosphate + 4 H+</td>
</tr>
<tr>
<td>cobD</td>
<td>116835</td>
<td></td>
<td>TIGR00380,pfam03186,PRK01209,COG1270</td>
<td>adenosylcobinate + (R)-1-amino-2-propanol O2-phosphate + ATP ↔ adenosyl-cobinamide phosphate + ADP + phosphate + H+</td>
</tr>
<tr>
<td>cobP</td>
<td>231831</td>
<td></td>
<td>pfam02283,COG2087,cd00544,PRK05800</td>
<td>adenosyl-cobinamide phosphate + GTP + H+ ↔ adenosylcobinate GDP + diphosphate</td>
</tr>
<tr>
<td>Protein</td>
<td>Accession</td>
<td>Source</td>
<td>Reaction</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>---------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>cobU_aerobe</td>
<td>231835</td>
<td>(Cameron et al, 1991)</td>
<td>β-nicotinate D-ribonucleotide + 5,6-dimethylbenzimidazole ↔ nicotinate + α-ribazole 5'-phosphate + H+</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td>α-ribazole 5'-phosphate + H2O ↔ α-ribazole + phosphate</td>
<td></td>
</tr>
<tr>
<td>cobV</td>
<td>71152737</td>
<td>(Cameron et al, 1991)</td>
<td>adenosylcobinamide-GDP + α-ribazole ↔ coenzyme B12 + GMP + H+</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 – Collection of sequences for the early Cobalt Insertion / Anaerobic pathway for B12 synthesis

<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Exp. Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>cysG</td>
<td>118137</td>
<td>(Wu et al., 1991; Stroupe et al., 2003)</td>
<td>TIGR01467,COG1648,TIGR01469, pfam10414,PRK10637,pfam01903, pfam13241,cd11642,pfam14824,COG0007</td>
<td>S-adenosyl-L-methionine + uroporphyrinogen-III &lt;=&gt; S-adenosyl-L-homocysteine + precorrin-1 + H+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S-adenosyl-L-methionine + precorrin-1 &lt;=&gt; S-adenosyl-L-homocysteine + precorrin-2</td>
</tr>
<tr>
<td>cbiK</td>
<td>543950</td>
<td>(Roth et al., 1993; Roessner et al., 1992; Raux et al., 1997; Schubert et al., 1999; Romão et al., 2011; Kaneko et al., 1995, 1996; Leech et al., 2003)</td>
<td>pfam06180,COG4822,cd03413,cd03412 PRK00923,pfam01903,COG2138,cd03416,cd03414</td>
<td>sirohydrochlorin + Co2+ &lt;=&gt; cobalt-sirohydrochlorin + 2 H+</td>
</tr>
<tr>
<td>cbiXL</td>
<td>48427998</td>
<td>(Roth et al., 1993; Roessner et al., 1992)</td>
<td>TIGR01467,cd11645,PRK05576</td>
<td>S-adenosyl-L-methionine + cobalt-sirohydrochlorin &lt;&lt;&lt; S-adenosyl-L-homocysteine + cobalt-factor III + H+</td>
</tr>
<tr>
<td>cbiL</td>
<td>20141258</td>
<td>(Roth et al., 1993; Roessner et al., 1992)</td>
<td>TIGR01467,cd11645,PRK05576</td>
<td>cobalt-precorrin-3 + S-adenosyl-L-methionine &lt;&lt;&lt; cobalt-precorrin-4 + S-adenosyl-L-homocysteine</td>
</tr>
<tr>
<td>cbiH</td>
<td>543948</td>
<td>(Roth et al., 1993; Roessner et al., 1992; Santander et al., 2006)</td>
<td>TIGR01466,PRK15478,cd11646,COG1010</td>
<td>S-adenosyl-L-methionine + cobalt-precorrin-4 &lt;=&gt; S-adenosyl-L-homocysteine + cobalt-precorrin-5A + H+</td>
</tr>
<tr>
<td>cbiF</td>
<td>61220314</td>
<td>(Roth et al., 1993; Roessner et al., 1992; Kajiwara et al., 2006)</td>
<td>TIGR01465,PRK15473,COG2875,cd11641</td>
<td>S-adenosyl-L-methionine + cobalt-precorrin-4 + S-adenosyl-L-homocysteine + cobalt-precorrin-5A + H+</td>
</tr>
<tr>
<td>cbiG</td>
<td>543947</td>
<td>(Roth et al., 1993; Roessner et al., 1992)</td>
<td>pfam01890,cd11760,cd11761, COG2073,PRK05788</td>
<td>cobalt-precorrin-5A + H2O &lt;=&gt; cobalt-precorrin-5B + acetaldehyde + H+</td>
</tr>
<tr>
<td>cbiD</td>
<td>543944</td>
<td>(Roth et al., 1993; Roessner et al., 2005)</td>
<td>TIGR00312,pfam01888,PRK00075,COG1903</td>
<td>cobalt-precorrin-5B + S-adenosyl-L-methionine &lt;=&gt; cobalt-precorrin-6A + S-adenosyl-L-homocysteine</td>
</tr>
<tr>
<td>cbiJ</td>
<td>543949</td>
<td>(Roth et al., 1993)</td>
<td>TIGR00315,cd02571,PRK08057,COG2099</td>
<td>cobalt-precorrin-6A + NAD+ &lt;&lt;&lt; cobalt-precorrin-6B + NAD+</td>
</tr>
<tr>
<td>cbiT</td>
<td>543957</td>
<td>(Roth et al., 1993; Roessner et al., 1992; Santander et al., 2006)</td>
<td>TIGR02469,PRK08287</td>
<td>cobalt-precorrin-6B + S-adenosyl-L-methionine + H+ &lt;=&gt; cobalt-precorrin-7 + S-adenosyl-L-homocysteine + CO2</td>
</tr>
<tr>
<td>cbiE</td>
<td>61220317</td>
<td>(Roth et al., 1993; Roessner et al., 1992; Santander et al., 2006)</td>
<td>TIGR02467,cd01980,cd11644, PRK05787,CDG2241</td>
<td>cobalt-precorrin-7 + S-adenosyl-L-methionine &lt;=&gt; cobalt-precorrin-8 + S-adenosyl-L-homocysteine</td>
</tr>
<tr>
<td>cbiC</td>
<td>543943</td>
<td>(Roth et al., 1993; Roessner et al., 1992)</td>
<td>pfam02570,COG2082,PRK08286,PRK05782</td>
<td>cobalt-precorrin-8 &lt;&lt;&lt; cobyrinate + H+</td>
</tr>
<tr>
<td>cbiA</td>
<td>20141224</td>
<td>(Roth et al., 1993; Richter-Dahlfors &amp; Andersson, 1992; Fresquet et al., 2004)</td>
<td>TIGR00379,pfam01856,PRK01077,COG1797</td>
<td>cobyribinate + 2 L-glutamine + 2 ATP + 2 H2O &lt;&lt;&lt; cobyribinate + 2 L-glutamate + 2 ADP + 2 phosphate + 2 H+</td>
</tr>
<tr>
<td>fldA</td>
<td>48428121</td>
<td>(Fonseca &amp; Escalante-Semerena, 2001)</td>
<td>COG0716,TIGR01752,PRK09267</td>
<td>2 cob(I)irurin a,c-diamide + FMN + 3 H+ &lt;&lt;&lt; 2 cob(I)irurin a,c-diamide + FMNH2</td>
</tr>
<tr>
<td>coba_anaerobe</td>
<td>399274</td>
<td>(Suh &amp; Escalante-Semerena, 1993, 1995; Escalante-Semerena et al, 1990; Fonseca et al, 2002; Bauer et al, 2001)</td>
<td>TIGR00708,TIGR00636, pfam02572,pfam01923</td>
<td>cob(I)irurin a,c-diamide + ATP &lt;&lt;&lt; adenosyl-cobyribinate a,c-diamide + PPPi</td>
</tr>
<tr>
<td>cbiP</td>
<td>543955</td>
<td>(Roth et al., 1993)</td>
<td>TIGR00313,pram01856,COG1492,PRK00784</td>
<td>adenosyl-cobyribinate a,c-diamide + 4 L-glutamine + 4 ATP + 4 H2O &lt;=&gt; 4 L-glutamate + adenosylcobyrate + 4 ADP + 4 phosphate + 4 H+</td>
</tr>
<tr>
<td>cbiB</td>
<td>543942</td>
<td>(Roth et al., 1993; Brushaber et al, 1998)</td>
<td>TIGR00380,pfam03186,COG1270,PRK01209</td>
<td>adenosylcobyrate + (R)-1-amino-2-propanol O2-phosphate + ATP &lt;&lt;&lt; adenosyl-cobinamide phosphate + ADP + phosphate + H+</td>
</tr>
</tbody>
</table>
Table 3.3 – Collection of sequences for synthesis of the lower ligand of cobalamin

<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>bluB</td>
<td>81634473</td>
<td>(Campbell et al., 2006; Taga et al., 2007)</td>
<td>cd02145,TIGR02476</td>
<td>FMNH2 + oxygen → 5,6-dimethylbenzimidazole + D-erythrose 4-phosphate + dialurate</td>
</tr>
<tr>
<td>bzaA</td>
<td>308741497</td>
<td>(Hazra et al., 2015)</td>
<td>PRK13352,pfam01964,COG0422,PRK09284,TIGR04386</td>
<td>5-amino-1-(5-phospho-β-D-ribo)imidazolyl + S-adenosyl-L-methionine + an reduced unknown electron acceptor → 5-hydroxy-benzimidazole + ammonium + formate + 5′-deoxyadenosine + L-methionine + an oxidized unknown electron acceptor + phosphate + 2 H+</td>
</tr>
<tr>
<td>bzaB</td>
<td>308741498</td>
<td>(Hazra et al., 2015)</td>
<td>TIGR04386,PRK13352,pfam01964,COG0422,PRK09284,PLN02444</td>
<td>5-amino-1-(5-phospho-β-D-ribo)imidazolyl + S-adenosyl-L-methionine + an reduced unknown electron acceptor → 5-hydroxy-benzimidazole + ammonium + formate + 5′-deoxyadenosine + L-methionine + an oxidized unknown electron acceptor + phosphate + 2 H+</td>
</tr>
<tr>
<td>bzaC</td>
<td>308741500</td>
<td>(Hazra et al., 2015)</td>
<td>cd02440,pfam12847,COG2890,pfam10050</td>
<td>5-hydroxy-benzimidazole + S-adenosyl-L-methionine → 5-methoxybenzimidazole + S-adenosyl-L-homocysteine + H+</td>
</tr>
<tr>
<td>bzaD</td>
<td>308741501</td>
<td>(Hazra et al., 2015)</td>
<td>TIGR04385,COG1032,pfam04055,TIGR04072</td>
<td>5-methoxybenzimidazole + S-adenosyl-L-methionine → 5-methoxy-6-methylbenzimidazole + S-adenosyl-L-homocysteine + H+</td>
</tr>
<tr>
<td>bzaE</td>
<td>308741502</td>
<td>(Hazra et al., 2015)</td>
<td>No conserved domains</td>
<td>5-methoxy-6-methylbenzimidazole → 5,6-dimethylbenzimidazole</td>
</tr>
</tbody>
</table>
### Table 3.4 – Collection of sequences for cobalamin dependent enzymes

<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>metH</td>
<td>34395941</td>
<td>(Old et al., 1990; Goulding et al., 1997)</td>
<td>TIGR02082,pfam02310,pfam02574, pfam02607,pfam00809, pfam02965</td>
<td>L-homocysteine + an N5-methyl-tetrahydrofolate → L-methionine + a tetrahydrofolate</td>
</tr>
<tr>
<td>nrdZ (rnr II)</td>
<td>74575572</td>
<td>(Dawes et al., 2003; Warner et al., 2007)</td>
<td>TIGR02504,TIGR02505 ,pfam08471,cd02888</td>
<td>a 2'-deoxyribonucleoside 5'-diphosphate + an oxidized thioredoxin + H2O = a ribonucleoside diphosphate + a reduced thioredoxin</td>
</tr>
<tr>
<td>nrdJ (rnr II)</td>
<td>499307400</td>
<td>(Gleason &amp; Olszewski, 2002)</td>
<td>TIGR02505,cd01676,cd01676</td>
<td>a 2'-deoxyribonucleoside 5'-diphosphate + an oxidized thioredoxin + H2O = a ribonucleoside diphosphate + a reduced thioredoxin</td>
</tr>
</tbody>
</table>

### Table 3.5 – Collection of sequences for cobalamin independent isoforms of the enzymes in Table 3.4

<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>metE</td>
<td>2851502</td>
<td>(González et al., 1992, 1996)</td>
<td>TIGR01371,TIGR01371,pfam08267,pfam01717, COG0620</td>
<td>L-homocysteine + N5-methyl-tetrahydropteroyl tri-L-glutamate → L-methionine + tetrahydropteroyl tri-L-glutamate</td>
</tr>
<tr>
<td>nrdAB (rnr I)</td>
<td>250730457014104</td>
<td>(Carlson et al., 1984; Nilsson et al., 1988; Nordlund et al., 1990; Uhlin &amp; Eklund, 1994)</td>
<td>cd01679,PRK09103,PHA02572,COG0209,TIGR02506 PRK09101,cd01049</td>
<td>an oxidized thioredoxin + a 2'-deoxyribonucleoside 5'-diphosphate + H2O ← a reduced thioredoxin + a ribonucleoside diphosphate</td>
</tr>
<tr>
<td>nrdEF (rnr I)</td>
<td>25073052507306</td>
<td>(Jordon et al., 1994, 1996)</td>
<td>TIGR04170,pfam08343,PRK08188,cd01679 COG0208,PRK13965,TIGR04171</td>
<td>dCDP + an oxidized NrdH glutaredoxin-like protein + H2O ← CDP + a reduced NrdH glutaredoxin-like protein</td>
</tr>
<tr>
<td>nrdDG (rnr I)</td>
<td>3439594871159601</td>
<td>(Sun et al., 1993, 1995)</td>
<td>TIGR02487,COG1328,PRK09263,cd01675,pfam13597,COG0602,TIGR02491,PRK11121</td>
<td>a reduced flavodoxin + a ribonucleoside triphosphate = an oxidized flavodoxin + a deoxyribonucleoside triphosphate + H2O</td>
</tr>
</tbody>
</table>

### Table 3.6 – Collection of sequences for biotin synthesis

<table>
<thead>
<tr>
<th>Gene name</th>
<th>NCBI GI</th>
<th>Reference</th>
<th>Conserved domains</th>
<th>Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioF</td>
<td>115010</td>
<td>(Lin et al., 2010)</td>
<td>TIGR00858,COG0156,PRK05958</td>
<td>Pimeloyl-[acyl-carrier protein] + L-alanine = 8-amino-7-oxononanoate + CO2 + holo-[acyl-carrier protein].</td>
</tr>
<tr>
<td>BioA</td>
<td>584840</td>
<td>(Lin et al., 2010)</td>
<td>TIGR00508,COG0161,PRK07986</td>
<td>S-adenosyl-L-methionine + 8-amino-7-oxononanoate = S-adenosyl-4-methylthio-2-oxobutanate + 7,8-diaminononanoate.</td>
</tr>
<tr>
<td>BioD</td>
<td>58484167462187</td>
<td>(Huang et al., 1995)</td>
<td>COG0132,PRK00090,TIGR00347,cd03109</td>
<td>Dethiobiotin + sulfur-(sulfur carrier) + 2 S-adenosyl-L-methionine + 2 reduced [2Fe-2S] ferredoxin = biotin + (sulfur carrier) + 2 L-methionine + 2 5'-deoxyadenosine + 2 oxidized [2Fe-2S] ferredoxin</td>
</tr>
<tr>
<td>Gene name</td>
<td>NCBI GI</td>
<td>Reference</td>
<td>Conserved domains</td>
<td>Reaction</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>--------------------------</td>
<td>------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>nadB</td>
<td>34395939</td>
<td>(Seifert et al., 1990)</td>
<td>COG0029,TIGR00551,PRK09077</td>
<td>L-aspartate + O2 = iminosuccinate + H2O2.</td>
</tr>
<tr>
<td>nadA</td>
<td>2507072</td>
<td>(Ollagnier-de Choudens et al., 2005)</td>
<td>pfam02445,TIGR00550,COG0379</td>
<td>Glycerone phosphate + iminosuccinate = pyridine-2,3-dicarboxylate + 2 H2O + phosphate.</td>
</tr>
<tr>
<td>nadC</td>
<td>2507263</td>
<td>(Hughes et al., 1993)</td>
<td>TIGR00078,COG0157,PRK09916</td>
<td>Beta-nicotinate D-ribonucleotide + diphosphate + CO2 = pyridine-2,3-dicarboxylate + 5-phospho-alpha-D-ribose 1-diphosphate.</td>
</tr>
<tr>
<td>KMO</td>
<td>75446355</td>
<td>(Crozier &amp; Moran, 2007)</td>
<td></td>
<td>L-kynurenine + NADPH + O2 = 3-hydroxy-L-kynurenine + NADP+ + H2O</td>
</tr>
</tbody>
</table>
| kynY      | 75345309 | (Momany et al., 2004)    | TIGR01814                          | L-kynurenine + H2O = anthranilate + L-alanine  
L-3-hydroxykynurenine + H2O = 3-hydroxyanthranilate + L-alanine |
| had       | 110278810| (Zhang et al., 2005)     | pfam06052,PRK13264,TIGR03037      | 3-hydroxyanthranilate + O2 = 2-amino-3-carboxyoxymonate semialdehyde.    |
| hpaB      | 426020938| (Liu et al., 2010)       | pfam11794,TIGR02309               | Anthranilate + FADH2 + O2 = 3-hydroxyanthranilate + FAD + H2O  
4-hydroxyphenylacetate + FADH2 + O2 = 3,4-dihydroxyphenylacetate + FAD + H2O |

Table 3.7 – Collection of sequences for niacin synthesis
3.3 Results

3.3.1 Predicting B12 synthesis phenotypes using comparative genomics
As this study is the first time an analysis of the B12 synthesis pathway has been attempted on this scale and across this diversity of bacteria, it was of interest to confirm that a comparative genomics approach was appropriate to accurately assign B12 biosynthesis capabilities. To do this, data on B12 synthesis in bacteria was gathered from literature and extended with data from members of my lab. Using this experimental dataset, it was possible to classify 43 bacterial species as either a ‘B12-producer’ or ‘not B12-producer’. The described computational pipeline was then used to predict B12 synthesis phenotypes for each of the 43 species and these results were then compared with the experimental classification (Table 3.8). As both B12 synthesis pathways include 20 enzymes, one of the biggest challenges in using a comparative genomics approach was to set an appropriate cut-off for the number of enzymes required for classification as a B12-producer, i.e. should a species with 17 out of 20 B12 synthesis enzymes be classified as B12-producer. Using the collected data on experimentally validated phenotypes the Rand accuracy of the predictions was tested, using a range of cut-offs for positive B12 synthesis assignment from 1 to 20 enzymes required.

\[
\text{Rand accuracy} = \frac{(TP + TN)}{(P + N)} = \frac{41}{43} = 0.953
\]

The data from this analysis is presented in Figure 3.3 and shows that the best correlation between the experimental data and computationally derived predictions is achieved when using a threshold where a bacterial species must have 15 genes or more from the biosynthesis pathway for it to be classified as a B12-producer. While lower thresholds also yielded similar accuracies, this is suspected to be influenced by the low number of non-B12-producers used in the study (8 out of 43). The most stringent cut-off was therefore selected, while maintaining a high accuracy.
Using this threshold of >70%, 41 out of 43 species were correctly classified as either ‘B12-producing’ or ‘non-B12-producing’ giving the method a Rand accuracy of 0.953. One of the species that was not correctly classified is Mycobacterium tuberculosis, which has been found to have all but cobF of the required genes for B12 synthesis (Young et al, 2015). It has never been cultured under conditions where presence of B12 has been detected (Warner et al, 2007; Savvi et al, 2008).

This comparison of the experimental data and predicted phenotypes gave enough confidence in the method to expand the analysis further.
Table 3.8 – Comparison of experimentally demonstrated and predicted phenotypes for 43 bacterial species. The species were computational prediction and experimental data are in alignment are assigned a “+”, while the species where this is not the case are assigned a “-“. From the comparison it is found that it is possible to correctly assign phenotypes in 41 out of 43 species giving the method a Rand accuracy of 0.953.

<table>
<thead>
<tr>
<th>Strain</th>
<th>TaxID</th>
<th>Pred. match (+ / -)</th>
<th>Experimental Reference</th>
<th>metE</th>
<th>metH</th>
<th>rnr I</th>
<th>rnr II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetobacterium woodii</td>
<td>33952</td>
<td>+</td>
<td>(Slusarchuk et al, 1988)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Aphanizomenon flos-aquae</td>
<td>1176</td>
<td>+</td>
<td>(Miyamoto et al, 2006)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Bacillus megaterium</td>
<td>1404</td>
<td>+</td>
<td>(Moore et al, 2013)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Brevibacterium panachium</td>
<td>497735</td>
<td>+</td>
<td>(Xie et al, 2013)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Brucella melitensis</td>
<td>29459</td>
<td>+</td>
<td>(Schroeder et al, 2009)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Burkholderia cepacia</td>
<td>292</td>
<td>+</td>
<td>(Xie et al, 2013)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Crocosphaera watsonii</td>
<td>263511</td>
<td>+</td>
<td>(Bonnet et al, 2010)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Desulfovibrio desulfuricans</td>
<td>876</td>
<td>+</td>
<td>(Choi &amp; Bartha, 1993)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dinoroseobacter shibae</td>
<td>215813</td>
<td>+</td>
<td>(Wagner-Döbler et al, 2010)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Eubacterium hallii</td>
<td>39488</td>
<td>+</td>
<td>(Engels et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Lactobacillus corynformis</td>
<td>1610</td>
<td>+</td>
<td>(Martin et al, 2005)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Lactobacillus reuteri</td>
<td>1598</td>
<td>+</td>
<td>(Santos et al, 2008)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Lactobacillus rossiae</td>
<td>231049</td>
<td>+</td>
<td>(De Angelis et al, 2014)</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Mesorhizobium loti</td>
<td>381</td>
<td>+</td>
<td>(Grant et al, 2014)</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Nitratireductor pacificus</td>
<td>1231180</td>
<td>+</td>
<td>(Payne et al, 2014)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Propionibacterium acnes</td>
<td>1747</td>
<td>+</td>
<td>(Kang et al, 2015)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Propionibacterium freudenreichii</td>
<td>1744</td>
<td>+</td>
<td>(Bykhovskii et al, 1998)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas denitrificans</td>
<td>43306</td>
<td>+</td>
<td>(Warren et al, 2002)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas putida</td>
<td>303</td>
<td>+</td>
<td>(Xie et al, 2013)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rhizobium leguminosarum</td>
<td>384</td>
<td>+</td>
<td>Our lab</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rhodobacter capsulatus</td>
<td>1061</td>
<td>+</td>
<td>(McGoldrick et al, 2005)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Rhodobacter sphaeroides</td>
<td>1063</td>
<td>+</td>
<td>(Erb et al, 2008)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Rhodospirillum rubrum</td>
<td>1085</td>
<td>+</td>
<td>(Gray &amp; Escalante-Semerena, 2007)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Ruegeria pomeroyi</td>
<td>89184</td>
<td>+</td>
<td>(Durham et al, 2014)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>28901</td>
<td>+</td>
<td>(Warren et al, 2002)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Sinorhizobium meliloti</td>
<td>382</td>
<td>+</td>
<td>(Campbell et al, 2006)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Streptomyces olivaceus</td>
<td>47716</td>
<td>-</td>
<td>(Bykhovskii et al, 1998)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Synechococcus sp. CC9311</td>
<td>64471</td>
<td>+</td>
<td>(Helliwell et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Synechococcus sp. WH 5701</td>
<td>69042</td>
<td>+</td>
<td>(Helliwell et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Synechococcus sp. WH 7803</td>
<td>32051</td>
<td>+</td>
<td>(Helliwell et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Synechococcus sp. WH 7805</td>
<td>59931</td>
<td>+</td>
<td>(Helliwell et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Synechococcus sp. WH 8102</td>
<td>84588</td>
<td>+</td>
<td>(Helliwell et al, 2016)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Synechocystis sp. PCC 6803</td>
<td>1148</td>
<td>+</td>
<td>Our lab</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Thermosiphio africanus</td>
<td>2421</td>
<td>+</td>
<td>(Swithers et al, 2012)</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Strain</td>
<td>TaxID</td>
<td>Pred. match (+/−)</td>
<td>Experimental Reference</td>
<td>metE</td>
<td>metH</td>
<td>mr class I</td>
<td>Rnr class II</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------</td>
<td>-------------------</td>
<td>------------------------</td>
<td>------</td>
<td>------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Bacillus clausii</td>
<td>79880</td>
<td>+</td>
<td>(Sudarsan et al., 2006)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Bacteroides thetaiaomicron</td>
<td>818</td>
<td>+</td>
<td>(Degnan et al., 2014b)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>562</td>
<td>+</td>
<td>Our lab</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Mycobacterium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tuberculosis</td>
<td>1773</td>
<td></td>
<td>(Warner et al., 2007; Savvi et al., 2008)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Myxococcus xanthus</td>
<td>34</td>
<td>+</td>
<td>(Bretscher &amp; Kaiser, 1978)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Prauserella rugosa</td>
<td>43354</td>
<td>+</td>
<td>(Bykhovskii et al., 1998)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Serratia marcescens</td>
<td>615</td>
<td>+</td>
<td>(Xie et al., 2013)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Synechococcus sp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCC 7002</td>
<td>32049</td>
<td>+</td>
<td>Our lab</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>2336</td>
<td>+</td>
<td>(Swithers et al., 2012)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
3.3.2 Prevalence and distribution of B12 synthesis in bacteria

After the experimental testing of the method, the analysis was expanded to include 8425 bacterial species, which is an order of magnitude more species than the largest previous study (540 species by Zhang et al, 2009). I found that 3601 species of bacteria have the required enzymes for B12 synthesis, which is similar to the 36.2% found by Zhang et al, (2009). To get a better sense of the distribution of the B12 synthesis trait, the individual species were grouped into their respective taxonomical orders and the frequency of the B12 synthesis in each order was calculated (Figure 3.4). This enables a fairer comparison between highly sequenced orders of bacteria such as Enterobacteriales (370 sequenced species) and less sequenced orders Prochlorales (11 sequenced species). On average between the orders, 37.2% of the species in each order have the genes required for B12 synthesis. From Figure 3.4 it can be observed that the distribution of the B12 synthesis trait is not uniform, but rather clustered together in certain groups. For example, the majority of the cyanobacteria (Stigonematales, Nostocales, Prochlorales, Oscillatoriales and Chroococcales) are B12-producers, while this trait is very rare in Lactobacillus (Firmicute), Legionellales (gamma-proteobacteria) and Flavobacteriales (Bacteroidetes).

These last three orders are unrelated, demonstrating the stochastic nature of the distribution of B12 biosynthesis.

Table 3.9 – Definitions of various categories of B12 metabolism found in bacteria and referenced in this study.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>B12-producer</td>
<td>Has more than 75% of either one of the two B12 synthesis pathways</td>
</tr>
<tr>
<td>Cobalamin-producer</td>
<td>Same as above but also has the enzymes required to produce DMB using either bluB or the bza operon</td>
</tr>
<tr>
<td>Pseudocobalamin-producer</td>
<td>Has the B12 synthesis pathway, but lack the ability to produce DMB</td>
</tr>
<tr>
<td>B12-utilizer</td>
<td>A species that have at least one B12-utilizing enzyme (either metH or rnr II)</td>
</tr>
<tr>
<td>B12-obligate</td>
<td>A species that have either metH or rnr II and not the corresponding metE and rnr I</td>
</tr>
<tr>
<td>B12-auxotroph</td>
<td>Same as above, but also lacking the B12 synthesis pathway and thus dependent on exogenous B12</td>
</tr>
</tbody>
</table>
3.3.3 Comparison of the early- and late cobalt insertion pathways

Two distinct versions of the cobalamin synthesis pathway exist, the early cobalt insertion pathway (also referred to as the anaerobic pathway) and the late cobalt insertion pathway (aerobic pathway), and the evolution of these two equally complex pathways remains an unsolved mystery even today. To investigate B12 biosynthetic capacity in the eubacterial kingdom further, each of the species capable of B12 synthesis was classified as either using the early- or late cobalt insertion pathway based on the presence of absence of the cobNST complex from the late cobalt insertion pathway, encoding the three subunits of cobaltochelatase, a Class I chelatase (Saha et al, 2009); in the anaerobic pathway this step is catalysed by a completely different enzyme CbiX (or CbiK), a Class II chelatase. Across all bacteria, 88.5% of the 3601 B12-producing species use the early cobalt insertion pathway, while the remaining 11.5% use the late cobalt insertion pathway. The distribution of the two pathways across bacterial orders was plotted to assess whether there is a taxonomical pattern in the distribution of the two pathways (Figure 3.5). It can be seen that the late cobalt insertion pathway is dominant only in the four alpha-proteobacterial groups, *Rhodobacterales, Rhizobiales, Rhodospirillales* and *Sphingomonadales*, while in all other bacterial orders, the anaerobic pathway predominates, and most
bacterial orders exclusively contain this synthesis pathway. An exception to this general trend is the bacterial order, *Burkholderiales* (part of the beta-proteobacteria), where 33.7% of the 448 species possess the early cobalt insertion pathway, while 7.6% have the late cobalt insertion pathway. I investigated this order further (Figure 3.6) and found that the genus *Burkholderia* contained an almost equal split between the early and late cobalt insertion pathways. As the presence of both pathways in one genus is quite unique, it merits further analysis which could potentially reveal species with both synthesis pathways.

Based on the results here it is not possible to determine which of the two pathways that is the oldest since the phylogenetic tree is unrooted. However, considering that the early cobalt insertion pathway is ~10x more abundant, found in all phyla and is dominant in ancient bacteria such as the cyanobacteria (Schirrmeister *et al.*, 2015), these results points towards it being basal, with the aerobic pathway arising at a later stage.

![Figure 3.5 – Distribution of the two B12 synthesis pathways. The early cobalt insertion pathway / anaerobic pathway (orange) and the late cobalt-insertion pathway / aerobic pathway (cyan). The tree is unrooted, and the coloured fields are used to separate different phyla.](image-url)
3.3.4 Variants of Vitamin B12

Several variants of B12 have been shown to exist in nature with the forms cobalamin and pseudocobalamin being among the most studied (Taga et al., 2007; Degnan et al., 2014b; Hazra et al., 2015). Given the reduced bioavailability of pseudocobalamin for humans (Stupperich & Nexø, 1991), knowledge of which bacteria produce cobalamin and pseudocobalamin may have a major impact on the fields of food nutrient and environmental sciences, but despite this no study has attempted to establish which of these two variants are most commonly produced by bacteria. Synthesis of cobalamin is thought to require available DMB, which forms the lower axial ligand of cobalamin. In pseudocobalamin
adenine is present instead of DMB. This study is based on the assumption that adenine is universally available and will be incorporated in the absence of DMB. In this study, the presence or absence of DMB synthesis was used to classify those species with the de novo B12 pathway as either capable of producing cobalamin or pseudocobalamin.

I wanted to quantify how large a fraction of B12-producing species was able to produce cobalamin and find the relative distribution of the two DMB synthesis pathways. From Figure 3.7 it can be seen how large a fraction of B12-producers that also are capable of DMB synthesis via either the bluB gene (purple) or the bza operon (green). Across all bacterial orders, 46.8% of B12-producing bacteria (1685 species) are also capable of DMB synthesis and thus likely to be cobalamin producers. The remaining 53.2% of B12-producers are likely to either produce pseudocobalamin, or another variant using available bases. It should be mentioned that some species such as *Listeria innocuato*, and *Salmonella typhimurium*, have been demonstrated to take up available DMB from the environment, and make cobalamin (Gray & Escalante-Semerena, 2010).

When studying the taxonomical distribution of cobalamin producing species, it is clear that again this trait is not evenly distributed among bacterial groups. Most B12-producers from the bacterial phyla *Firmicutes* lack the ability to produce cobalamin, with the exception of the order *Thermoanaerobacterales*, which is a group of thermophilic and anaerobic bacteria. The reason for this apparent requirement or preference for cobalamin is currently not known, but B12 has been linked to ethanolamine utilisation in *Thermoanaerobacter sp. X514* (Lin *et al.*, 2011), and it is conceivable that the ethanolamine deaminase might only function with cobalamin and not pseudocobalamin (or other variant) as a co-factor. The prevalence of the newly discovered bza operon, compared to the bluB gene, was calculated, to gain further understanding into the evolution and distribution of these two pathways. I found that 96.7% of the species capable of cobalamin synthesis encode bluB, while only 3.3% utilise the bza operon. The bza operon also seem to be mainly confined to the *delta-proteobacteria* phylum, with only a few examples outside this phylum. The limited taxonomical spread of the bza operon and its low frequency compared to bluB suggests that this pathway arose...
at later stage in evolution than the bluB pathway, but further studies are needed to verify this.

Figure 3.7 – Frequency of DMB synthesis among B12-producers from the various phylogenetic orders. Species with the bluB gene are shown in purple and species with the anaerobic (bzs operon) DMB synthesis pathway are shown in green. E.g. 100% of the B12-producers from Prochlorales produce pseudocobalamin, while 90% of the Rhizobiales produce cobalamin. The tree is unrooted, and the coloured fields are used to separate different phyla.

### 3.3.5 B12 synthesis prevalence in selected environments

Quantifying B12 synthesis prevalence in a taxonomical context revealed that this trait is not distributed equally among all bacterial groups (Figure 3.5). As B12 is a limiting nutrient in certain environments (Bertrand et al., 2007; Sañudo-Wilhelmy et al., 2012), it is interesting to analyse B12 synthesis frequency in an environmental context and see if this trait is equally common in different environments. Using data from GOLD (Reddy et al., 2015) on natural habitats, the environmental context was available for 1950 of the 8425 studied species. The frequency of B12 synthesis (early and late cobalt insertion pathways) was calculated for selected environments (Figure 3.8). Surprisingly, it was found that in most of the environments analysed, the frequency of B12 synthesising bacteria was similar to the average across all bacteria (37.2%), with only the rhizoplane (associated with plant roots in the soil) and thermal springs environments differing by more than 10 percentage points from this average. It is also worth noting that environment with the lowest frequency of B12-producers, ‘Food production’ (32.4%), still contains a significant number of B12-producers. ‘Food production’ refers to engineered communities used in dairy production (Ivanova et al., 2010). When comparing the frequency of the two pathways, it can be seen that both
pathways are found in eight out of nine environments, but that the early cobalt insertion pathway is dominant in all of them. The only environment where the late cobalt insertion pathway is present in more than 12% of the species is the rhizoplane, which has a large fraction of alpha-proteobacteria, such as Rhizobiales and Rhodobacteriales.

Further to this, the frequency of species that are predicted to be able to synthesise cobalamin (green) or not (grey) in selected environments is plotted in Figure 3.9. Both traits can be identified in all nine environments, but the relative frequency of two phenotypes varies greatly between them. Among the species from the digestive system and thermal springs, only a small fraction is capable of cobalamin synthesis, which points towards a preference for pseudocobalamin among most bacteria in these environments. In contrast to this, more than 90% of the B12-producing bacteria from the rhizoplane are capable of cobalamin synthesis and thus appear to have a preference for this variant. It is however unclear if this preference of a B12 variant in the mentioned environments points towards a biochemical advantage of one of the variants in certain environments or if it is simply a result of the stochastic nature of microbial environments and evolution.
Figure 3.9 – Fraction of species predicted to produce pseudocobalamin (grey) and cobalamin (green) in selected environments. The average frequency of B12 synthesis across all species is marked by the dotted line. The number of species analysed from each environment is displayed in square brackets.

### 3.3.6 B12 usage and auxotrophy in bacteria

In addition to analysing B12 synthesis capabilities in bacteria, B12 usage was also studied. Although there are over 20 B12-dependent enzymes in bacteria, I focused on the two major ones that are universal, and also are found in eukaryotes, namely B12-dependent methionine synthase (MetH), methyl- and type II ribonucleotide reductase (RNR II), since these have B12-independent variants (MetE and RNR I) respectively. When quantifying the presence of these different enzymes, 98.3% of the 8425 analysed species possess at least one enzyme (MetH or RNR II) that utilises B12 as a co-factor. This demonstrates that B12 is a widely used co-factor in bacteria, similar to what is observed in mammals (Froese & Gravel, 2010), but in contrast to the B12-free metabolism of plants (Smith et al, 2007) and fungi (Suliman et al, 2005). No noticeable patterns emerged when analysing B12 usage in a taxonomical context. On the other hand, when comparing the prevalence of the B12-dependent isoforms to the B12-independent isoforms of the two selected enzymes, it was found that having only the B12-dependent isoform is 5x more common than having only the B12-independent isoform for both of the investigated enzymes (Figure 3.10). This suggests that while it is a disadvantage to depend on a complex co-factor like B12, the metabolic advantage of the B12-dependent isoforms of the enzymes outweigh this ‘cost’.

![Figure 3.9](image_url)
Under the assumption that synthesis of methionine (by methionine synthase) and nucleotide synthesis (by ribonucleotide reductase) are critical process in all living cells, it is possible to calculate the fraction of bacteria that are reliant on external sources of B12. B12-auxotrophs were identified by finding species that do not have the B12 synthesis pathway and at the same time have an absolute requirement for B12 for the metabolism, e.g. having the metH gene, but not metE gene (Figure 3.11). When averaging the frequency of B12 auxotrophy from each order, it is found that 24.6% of bacteria are predicted to be dependent on exogenous B12 or an intermediate from the synthesis pathway. When looking across the different orders, it is clear that, like B12 synthesis, this is a trait that varies between the different taxonomical groups. In groups such as Myxococcales, Flavobacteriales and Thermotogales, more than 80% are dependent on external supplies of B12 or intermediates, while in Desulfovibrionales, Enterobacteriales and Prochlorales less than 5% are predicted to B12-dependent. Upon looking at Myxococcales (a group of delta-proteobacteria) it was found that found that 86% (25/29 species) are predicted to be B12-auxotrophs. Myxococcales (Myxobacteria) is a group of bacteria that can be found in soil and are known for having genomes larger than 9Mb, which makes them among the largest bacterial genomes discovered (Chen et al, 2016).

The finding that these bacteria are largely B12-dependent was initially surprising, considering that much smaller organisms such as Prochlorococcus marinus (1.75 Mb) encode a full B12 synthesis pathway (Bonnet et al, 2010). This led me to
undertake a more detailed analysis of genome size and presence/absence of B12 synthesis pathway. When binning species according to the number of predicted unique proteins (used as proxy for genome size) and then calculating the fraction of species that are capable of B12 synthesis in each bin, it is found that B12 synthesis is in 32.4% of species with 2-3K predicted proteins, while it is found in 83.7% of species with 9-10K predicted proteins (Figure 3.12). Based on this analysis it appears that there is a correlation between genome size and the frequency of the B12 synthesis pathway, although with exceptions such as Prochlorococcus. When this apparent correlation between genome size and presence of B12 synthesis pathway is taken into account, the rarity of the B12 synthesis pathway in Myxococcales is perhaps surprising considering the 9 Mb genomes in this group. The reason for this is not known, but one possible explanation is that the predatory behaviour of Myxobacteria means they could cover their B12 need from prey bacteria (Goldman et al, 2006; Berleman & Kirby, 2009). Another bacterial order with a high frequency (12/13 species) of apparent B12 auxotrophy is Thermotogales. This group of thermophilic bacteria, which are thought to be among the oldest lineages of bacteria (Butzin et al, 2013), have developed mechanisms to compensate for their B12-dependence. T. litiingae has been found to take up cobinamides and convert these to bio-active B12 variants as well as regular B12 uptake, using the btuCDF system, to cover its B12 requirements (Butzin et al, 2013).
B12 auxotrophy was also analysed in an environmental context using the same method and data as for the analysis of B12 synthesis. I found that the frequency of B12 auxotrophy is close to overall average in five of the investigated nine environments (Figure 3.13). Among the bacterial species involved in food production (4.1%) and associated with skin (12.9%), this trait is quite rare. At
the other end of the spectrum, B12 auxotrophy among bacteria from thermal springs is more common, where 37.5% of the bacteria are predicted dependent on B12, compared with 24.6% across all bacteria. Going more into detail with the bacteria from ‘thermal springs’ it is found that as 37.5% are B12-auxotrophs and 57.5% capable of B12 synthesis, meaning that 95% of the species from this environment either depend on or produce this co-factor (compared to 61.8% for all groups). This could imply this trait is particularly important in this harsh environment and that B12-utilising enzymes have advantages over the B12 independent counterparts. It has been found that B12-producing bacteria can enhance the thermal tolerance of the green algae C. reinhardtii (Xie et al, 2013) due to the increased thermal stability of METH (utilises B12) over METE (B12 independent). Potentially, this is the reason why there is a very high prevalence of B12-utilisation in this particular environment, but further studies will be needed to verify this.

Figure 3.13 – Frequency of B12 auxotrophy in selected environments. The average across all species is marked by the dotted line. The number of species analysed from each environment is displayed in square brackets. E.g. of the 222 fresh water bacteria studied, 22% were found to be B12-auxotrophs.

3.3.7 Niacin, biotin and B12 synthesis in algae-associated bacteria

Croft et al, (2005) showed that several eukaryotic algae are dependent on exogenous B12 and further to this it has been found that eukaryotic algae appear to have a preference for cobalamin over pseudocobalamin (Helliwell et al, 2016). The background levels of B12 in marine environments are lower than what is required by B12 dependent algae such as Ostreococcus tauri, and have
been found to be a limiting factor for algal productivity in marine environments (Bertrand et al, 2007, 2015). While it is known that bacteria can form communities with B12-dependent algae and supply them with B12, it is not known which are likely to provide the B12-requirement of algae in oceans (Kazamia et al, 2012b; Grant et al, 2014; Durham et al, 2014). A list of cobalamin producing bacteria had already been created (Data summarised in Section 3.3.4), the next step in the analysis was to identify bacteria that were often found together with algae and cross reference the two lists. By combining the results from eight environmental studies of algae and their associated bacteria, it was possible to identify two bacterial genera that were found most frequently in the phycosphere of several algal species and in multiple environments (Table 3.10). The species from the two alpha-proteobacterial genera Roseobacter (found in 5/8 studies) and Sulfitobacter (found in 7/8 studies) were analysed further and predictions were made regarding their ability to produce B12 and DMB (Table 3.11). It was found that 17/18 of the sequenced species from two genera were predicted to be able to produce B12 and possessed the DMB synthesis pathway and therefore likely to produce cobalamin, which can support B12 dependent eukaryotic algae (Helliwell et al, 2016).
Table 3.10 – Collection of studies of algae and bacterial genera from their phycosphere from various marine environments. From each study, the identified species and algae were extracted.

<table>
<thead>
<tr>
<th>Environment sampled</th>
<th>Algae</th>
<th>Bacterial genera</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pacific and Atlantic Ocean</td>
<td>Pseudo-nitzschia multiseries</td>
<td>Roseobacter, Sulfitobacter, Hyphomonas, Marinobacter, Limnobacter, Croceibacter</td>
<td>Amin et al., 2015</td>
</tr>
<tr>
<td>Mexican Gulf and Atlantic Ocean</td>
<td>Dylidium brightwellii, Thalassiosira weissogii, Asterionella glacialis, Chaetoceros socialis, Leptocylindrus danicus, Coscinodiscus sp.</td>
<td>Cytophaga, Flavobacterium, Bacteroides, Ruegeria, Sulfitobacter, Roseobacter, Erythrobacter</td>
<td>Schäfer et al., 2002</td>
</tr>
<tr>
<td>Arctic Ice</td>
<td>Amphiprora kufferathii</td>
<td>Sulfitobacter, Colwellia, Pibocella</td>
<td>Hünken et al., 2008</td>
</tr>
<tr>
<td>Northern Atlantic Ocean</td>
<td>Pseudo-nitzschia multiseries</td>
<td>Sulfitobacter, Labrenzia, Paracoccus, Novosphingobium, Pseudoalteromonas, Cellvibrio, Reichenbachia</td>
<td>Kaczmarska et al., 2005</td>
</tr>
<tr>
<td>North Sea</td>
<td>Guinardia delicatula, Pseudo-nitzschia pungens, Thalassiosira rotula, Skeletonema costatum, Ceratium horridum, Akashiwo sanguinea</td>
<td>Roseobacter, Sulfitobacter</td>
<td>Sapp et al., 2007</td>
</tr>
<tr>
<td>North Sea</td>
<td>Thalassiosira rotula, Skeletonema costatum</td>
<td>Sulfitobacter, Roseobacter, Thalassobius, Octadecabacter</td>
<td>Grossart et al., 2005</td>
</tr>
<tr>
<td>Pacific Ocean</td>
<td>Thalassiosira sp</td>
<td>Rimicaris, Lewinella, Ochrosphaera, Cytophaga, Prochlorococcus, Cyanobacteria, Roseobacter, Alteromonas</td>
<td>Riemann et al., 2000</td>
</tr>
</tbody>
</table>
This led to further investigation of these two bacterial genera, and from other studies it was found that growth of the cobalamin dependent marine diatom *Pseudo-nitzschia multiseries* was enhanced when the diatom is co-cultured with bacteria from the *Sulfitobacter* genus (Tang et al., 2010; Amin et al., 2015). It was also found that *Dinoroseobacter shibae*, which is closely related to *Sulfitobacter* and *Roseobacter*, is capable of supplying B12 to diatoms in a laboratory setting (Wagner-Döbler et al, 2010). Other previous studies had shown that the two *Roseobacter* species, *R. litoralis* and *R. denitrificans* require an exogenous source of niacin and biotin for growth and that *P. multiseries* is a demonstrated biotin-producer (Shiba, 1991; Tang et al, 2010). Together these observations led to a hypothesis that *Sulfitobacter* and *Roseobacter* and certain cobalamin dependent marine algae form a symbiotic relationship revolving around a two-way exchange of vitamin B3, B7 and B12 (Figure 3.14). It was investigated whether B3 and B7 auxotrophy is likely to be confined to *R. litoralis* and *R. denitrificans* or a general phenotype across the *Sulfitobacter* and *Roseobacter*. This was done using the same method applied for identification of presence or absence of B12 synthesis and the results are shown in Table 3.11. It was found that 16/18 species are B7 auxotrophs and that 13/18 species are B3 auxotrophs. The computational results was in agreement with the experimental data from Shiba (1991) on *R. litoralis* and *R. denitrificans*.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Sulfitobacter donghicola</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter geojensis</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter guttiformis</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter mediterraneus</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter noctilucae</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter noctilucicola</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter pseudonitzschiae</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><em>Sulfitobacter sp. 20_GPM-1509m</em></td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><em>Sulfitobacter sp. CB2047</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Sulfitobacter sp. EE</em></td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>√</td>
</tr>
<tr>
<td><em>Sulfitobacter sp. NAS-14.1</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter denitrificans</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter litoralis</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter sp. AzwK-3b</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter sp. CCS2</em></td>
<td>√</td>
<td>√</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter sp. GAI101</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Roseobacter sp. MED193</em></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><em>Roseobacter sp. SK209-2-6</em></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Coincidentally, in our laboratory a model system comprising the marine green alga *Osterococcus tauri* and the marine *alpha*-proteobacteria *Dinoroseobacter shibae* had been established (MB Cooper, PhD thesis 2014). This system enabled us to test the hypothesis (Figure 3.14). Figure 3.15a shows that *O. tauri* requires cobalamin and thiamine for growth, while *D. shibae* is depend on biotin, para-aminobenzoic acid and niacin for growth. Figure 3.15b and Figure 3.15c show when co-culturing the two organisms in the absence of these critical nutrients, growth is restored although at a lower level than when supplemented with vitamins. While the experiment was performed using related species, rather than *P. multiseries* and a *Sulfitobacter/Roseobacter* species, it shows as proof of principle that two-way B vitamin exchanges between a marine alga and bacterium is possible.

Figure 3.14 – Hypothesis of bi-directional B vitamin exchange between the marine diatom *P. multiseries* and bacteria from the genera *Sulfitobacter* and *Roseobacter*. It is hypothesised that ocean algae such as *P. multiseries* and *Sulfitobacter/Roseobacter* species form mutualistic relationships where B3 and B7 are exchanged for B12. It is likely that other metabolites are also exchanged.
Figure 3.15 – Experimental validation of two-way B vitamin exchange between the marine alga Ostereococcus tauri and the marine bacterium Dinoroseobacter shibae. A) Growth Assay in microtiter plate to validate vitamin auxotrophies for Ostereococcus tauri and Dinoroseobacter shibae. Bacterial cultures were also supplemented with 1% glucose. B) Growth of O. tauri in axenic (grey) and co-cultures with D. shibae (white) with and without various B vitamins. C) Growth of D. shibae in axenic (grey) and co-cultures with O. tauri with and without various B Vitamins. Figures from Cooper et al, (in preparation)
3.4 Discussion

3.4.1 Applying comparative genomics to identify B12 producing bacteria.

This study was started with the aim of identifying bacterial species capable of producing B12, but to do this a method first had to be developed. Based on experience and previous smaller studies, a comparative genomics method was selected. This method relied on homology searches of translated genomes from sequenced bacteria and representative sequences from the enzymes required for B12 biosynthesis. A similar version of this method has been applied in other studies, but either with a smaller pathway or a smaller set of studied species. Examples include a study of six selenium utilizing enzymes in several thousand species (Peng et al., 2016) or in a study on B12 synthesis enzymes but restricted to 313 species found in the human gut (Degnan et al., 2014a). As the goal of this study involved searching for >50 enzymes in >8000 species, I first had to demonstrate that the method was appropriate for a study of this magnitude and diversity. This was approached by collecting experimental data on species either with or without a demonstrated B12 synthesis pathway and then correlating computationally assigned phenotypes with the experimental data to identify the optimal parameters for the homology search. By validating the homology search, the confidence in the results increased dramatically and pointed towards comparative genomics as a suitable method to study long pathways across a very large range and diversity of bacteria. The parameters and search strategy developed here can easily be used as a basis for other studies, but it will still be advantageous to perform a similar comparison of experimental and predicted phenotypes for each new study. As the experimental and predicted phenotypes matched at a satisfactory level for this study, further exploration of search thresholds was not done. However, parameters such as e-values, alignment coverage between query and target sequences, number of matching domains etc. could be explored further and used to optimized correlation between experimental and predicted phenotypes.

A further refinement of the method would be to incorporate the genomic location of individual genes from the synthesis pathway. The genes required for B12 synthesis are found in operons scattered around the chromosome. Nonetheless, this would allow the analysis to be expanded using genome location of the
individual genes from the synthesis pathway. Additional, it has been shown that these operons are typically found in proximity to a B12 binding box that serves to regulate expression of the synthesis genes (Rodionov et al, 2002). The study by Rodionov et al, (2002) analysed the B12 synthesis in 87 species, while this study attempts to analyse ~100x more species. This difference in scale represents the main challenge in carrying out an operon structure analysis, as the study by Rodionov et al, (2002) also showed that the operon structure varies between species, so it is a major challenge to define a set of ‘rules’ that could be incorporated into the computational framework. Another expansion of the search strategy could be to include multiple query sequences in the homology search. This strategy was used by Degnan et al, (2014a) in their study of 313 bacterial species from the human gut. While this method could give more precise results, the comparison of the experimental and computational data in my study shows that using only one query sequence per enzyme in the synthesis pathways yields results with an acceptable level of accuracy, and thus any additional use of query sequences is redundant.

As the price of sequencing has been dropping over the past decade, the number of newly sequencing public bacterial genomes has been rising quickly (Stein, 2010). This requires any analysis pipeline to be scalable and compatible with the large stream of new genomes made available. The computational pipeline build for this project is developed so that given a new set of genomes in a text format, it is a simple matter to include these in the current analysis. If the efforts in sequencing of bacteria continues, then there will be an increased demand on the computational system behind the analysis. The limiting factor in the computational part of analysis is the available number of CPUs. To address this the computational framework was built so that it can be scaled according to the available computational resources, being able to use multiple CPUs and thus scaled to any number of CPUs available on the system. Enabling multiple CPUs led to this analysis being complete in ~4 days of running time, as opposed to ~160 days that it would take if using a single CPU. It should be noted that 160 days of running time is unlikely to be possible as few computer systems can operate continuously for 160 days. Enabling multiple CPUs for data processing was therefore essential to make a project of this magnitude possible.
3.4.2 *B12 synthesis, two pathway ways and two variants.*

The analysis of B12 synthesis across 8425 bacterial species found that 37.2% of bacteria are capable of B12 synthesis. This was very similar to the 36.2% found in study by Zhang *et al.*, (2009) that also sought to identify presence or absence of B12 synthesis capabilities across all bacteria (540 species studied). It was perhaps surprising that the estimated frequency of B12 synthesis in bacteria were so similar considering that this study investigated more than ten times as many species and thus represented a much broader range of species. On the other hand, perhaps this is a true reflection of the proportion of B12-synthesising bacteria. As more genomes are added to public genome databases and these genomes are analysed for presence B12 synthesis enzymes, it will be of interest to follow the estimated frequency of B12 synthesis and see if it remains in the range of 35-40% or if changes from this.

The taxonomical distribution of the B12 synthesis showed that this trait is found in most branches of the bacterial kingdom, but at different frequencies. One of the phyla where B12 synthesis is present in most species is in cyanobacteria. Cyanobacteria are thought to be among the oldest groups of bacteria and have been estimated to be as old as 3.5 billion years. They are thought to be responsible for ‘the great oxygenation event’ (Schirrmeister *et al.*, 2015), that changed the environment from an anaerobic state to an aerobic. As oxygen was only available in very limited quantities at the time this group evolved, it would appear obvious that cyanobacteria utilised the anaerobic synthesis pathway for B12 synthesis, where ring contraction occurs without the need for molecular oxygen. Along a similar line of thought, in a collaborative project in the laboratory, species of the abundant marine cyanobacterium genus *Synechococcus* was shown to make the pseudocobalamin variant of B12 and does not appear to have the *BluB* gene. This work was published as Helliwell *et al.*, (2016). B12 is thought to pre-date living organisms and it has been shown that it can be formed through non-enzymatic reactions (Hodgson & Ponnamperuma, 1968). The near omnipresence of the B12 synthesis pathway in cyanobacteria suggest that the synthesis pathway was present in the ancestral line of these bacteria, and thus places B12 as an ancient molecule. The observation that *Thermotogales*, which has been found to be the most ancient line of bacteria (Butzin *et al.*, 2013), are
dependent on exogenous supply of the vitamin points towards B12 usage among bacteria being a very old trait.

### 3.4.3 Studying B12 uptake using comparative genomics

One aspect of B12 usage that was not studied in this thesis is uptake of B12 from the environment surrounding a given bacteria. The \textit{btuCDF} uptake system was identified in \textit{E. coli} and has been described in high detail (Locher \textit{et al}, 2002; Borths \textit{et al}, 2005). In relation to this study the next logical step would be to attempt to identify the presence of the \textit{btuCDF} system in all bacterial species, and then to correlate B12 uptake with B12 auxotrophy. If there were species identified without a B12 uptake system and with B12 auxotrophy, they could potentially have an alternative uptake system currently not identified.

However, one of the main challenges in doing this is that identifying the function of transporters based on the sequence alone is highly challenging. This is due to the high degree of similarity in amino acid sequence between transporters despite them transporting very different metabolites (Genee \textit{et al}, 2016). In a study by Zhang \textit{et al}, (2009) the conclusion was reached that it is not possible to distinguish the \textit{btuCDF} system transporter from heme transporters simply based on sequence alone. One solution to this probably could be to look for a B12-binding box in close proximity to the putative \textit{btuCDF} genes to identify regulatory elements. This could be used to confirm or reject and potential \textit{btuCDF} candidates.

### 3.4.4 Expanding the study of algae associated bacteria and B-vitamin auxotrophy

Based on the observation that \textit{D. shibae} and \textit{O. tauri} can exchange B-vitamins in a bilateral manner it seemed obvious to study if this was likely to be happening only with these two organisms or if this was a general phenomenon. From the previously published studies, it was identified that \textit{Sulfitobacter} and \textit{Roseobacter} species are often found in the phycosphere of algae in marine environments. The study presented here encompasses data from nine environmental locations, but the search was not exhaustive. It would therefore be beneficial to expand the number of studies included to get stronger evidence for which bacteria are found in most phycospheres. Another potential expansion is to investigate experimental
data on niacin and biotin auxotrophy in *Roseobacter* and *Sulfitobacter*. As there was only data available for two species, the next step would be to carry out simple experiments demonstrating the apparent auxotrophies in more of the sequenced bacteria in these orders that are associated with algae.
4 Transcriptomics study of a green alga co-cultured with a bacterium for exchange of B12

Preface

The work described in this chapter is a result of a collaboration with multiple partners. Vaibhav Bhardwaj (formerly Department of Plant Sciences, University of Cambridge) and Dr Katherine Helliwell (formerly Department of Plant Sciences, University of Cambridge) grew the cultures used in the experiments and extracted RNA from them. Construction of the libraries of cDNA prepared from the RNA samples, and then sequencing of these libraries was carried out at the Cambridge Genomic Services, in the Department of Pathology, University of Cambridge by Dr Edward Farnell. Dr Krys Kelly (Department of Plant Sciences, University of Cambridge) and Chris Boursnell (Department of Plant Sciences, University of Cambridge) provided advice on the bioinformatics part of the study.

It should be noted that the work described in this chapter complements a proteomics study performed using the same model system of M. loti and L. rostrata. I co-authored the article describing this study and it has been accepted for publication in New Phytologist as Helliwell et al, (2017) and is included as Appendix 7.4.
4.1 Introduction

The photosynthetic green alga *L. rostrata* and the B12-producing soil bacterium *M. loti* have been demonstrated to form stable co-cultures when grown together (Kazamia *et al.*, 2012b; Grant *et al.*, 2014). While this community has been studied in terms of cell numbers and nutrient levels, the molecular mechanisms underpinning this mutualistic relationship are not fully understood. In a proteomics study of this community that I co-authored, the protein expression in *L. rostrata* was quantified from an axenic culture supplied with B12 and a co-culture of *L. rostrata* and *M. loti* (Helliwell *et al.*, 2017). From the experiment, 588 proteins were identified, and the analysis of these proteins showed expression of proteins related to amino acids synthesis was higher in co-cultures. Another central finding was that proteins related to photosynthesis were expressed at lower levels in co-cultures. This observation was validated by measuring electron transfer rate as a proxy for photosynthesis capacity and was found to be lower in co-cultures.

To expand the understanding of the interaction between these two organisms, a transcriptomics experiment was carried out with the aim of identifying further changes in gene expression between mono-cultures and co-cultures. An RNA-seq based approach was selected due to it being successfully applied in previous studies involving algae and bacteria co-cultures. Recent examples of include studies of interactions between diatoms and alpha-proteobacteria: *Thalassiosira pseudonana* with *Roseobacter pomeroyi* (Durham *et al.*, 2014), and *Pseudo-nitzschia multiseries* with *Sulfitobacter* (Amin *et al.*, 2015). Both studies are described in detail in Chapter 1.

RNA-seq was also chosen as there is no genome sequence available for *L. rostrata*, so performing a targeted qPCR study of selected transcripts would be challenging. On the other hand, without a reference genome, a de novo assembly process had to be undertaken to construct a representative of the *L. rostrata* transcriptome. The fundamental challenge in this field is to construct trustworthy longer sequences that represent individual transcripts from short sequence reads. As these longer sequences are only candidate transcripts, but not verified
they are referred to as contigs. The collection of contigs is referred to as an assembly rather than a transcriptome.

RNA-seq studies in algal species without a reference genome have been performed successfully previously, including for *Chlorella* sp. “SAG-211-18” (Mansfeldt *et al.*, 2016). They studied the *Chlorella* transcriptome under nutrient replete and deplete conditions to identify changes in gene expression. A central observation from the study was that when switching from nutrient replete to deplete conditions, transcripts encoding enzymes related to starch and lipid synthesis are upregulated, while transcripts encoding enzymes related to photosynthesis are down regulated. The phytoflagellate *Euglena gracilis* has been the subject of two transcriptome-based studies recently. Yoshida *et al.*, (2016) sequenced cultures of *E. gracilis* under aerobic and anaerobic conditions and constructed a *de novo* assembly of the transcriptome. The authors identified a set of candidate genes related to synthesis of wax esters but concluded that regulation of this process was more likely to be post-translational. In a similar study, O’Neill *et al.*, (2015) also constructed a representation of the *E. gracilis* transcriptome using RNA-seq. The generated assembly was used to place *E. gracilis* in a phylogenetic context and the authors determined that *E. gracilis* contains genes with homologs in mammals, plants and bacteria.

Building a representative transcriptome by *de novo assembly* for algal species is becoming more common these days, but still provides a significant challenge. This study is to my knowledge also the first time a transcriptome study of a co-culture has been undertaken with an un-sequenced alga. The data from this study will also provide the first opportunity to place *L. rostrata* in a phylogenetic context using large quantities of nucleotide data.

The study had the following aims:

1) To demonstrate that it is possible to perform a transcriptomics experiment of a synthetic co-culture consisting of a bacterium and an alga, including a conducting a *de-novo* transcriptome assembly from a co-culture.

2) To characterise changes in the metabolism of the organisms based on changes in transcript levels
3) To uncover specific genes that could have a central role in the mutualistic interaction between *L. rostrata* and *M. loti*

### 4.2 Methods

#### 4.2.1 Cultivation and harvesting of cells

Cultures used for the RNA-seq experiment were grown by Vaibhav Bhardwaj using the following protocol.

Axenic cultures of *Lobomonas rostrata* were maintained on TAP agar + B12 slopes at 21 °C. Material for the RNAseq experiment was obtained by the following method. Single colonies were selected and cultured in 10 mL of heterotrophic medium (TP*(Kazamia *et al*, 2012b) in a shaking incubator at 25 °C, 120 rpm under white light of 100 µM m⁻² s⁻¹. After 14 days, cells from this preculture were inoculated into 50 mL of TP*. To ensure equal cell numbers were inoculated, cells counts were carried out using a Beckman Coulter counter and an inoculation volume was chosen such that an initial concentration of 50 000 cells mL⁻¹ was used. Flasks were then grown in a shaking incubator 16 h light 8 h dark cycles, 120 rpm and a temperature of 25 °C. Cells were harvested after 8 days.

The strain *Mesorhizobium loti* MAFF303099 was used for this study and cultured on TY plates from cryopreserved stocks. Plates were incubated at 28 °C for 72 h and then single colonies were picked and inoculated into 5 mL of TP* supplemented with glycerol. The cultures were incubated in a shaking incubator at 120 rpm and at 28 °C for 72 h. The OD of the cultures was measured and readjusted to an OD of 0.1 at 595 nm. A 100-fold dilution was then carried out thus giving an inoculum concentration of approximately 1 X 10⁵ bacteria mL⁻¹. The cultures were then grown in 16h light 8 h dark cycles. Cells were harvested at an OD between 0.3 and 0.5 at a wavelength of 595 nm.

Co-cultures were set up using individual pre-cultures of the algae and bacteria prepared as above. The precultures were grown for 21 days and combined into 50 mL flasks. After two subcultures, a stable co-culture was inoculated into fresh TP* at an initial concentration of 50 000 algal cells mL⁻¹. The flasks were grown
in 16h light, 8 h dark conditions with shaking, and cells were harvested after 8 days.

All cultures were harvested by centrifuging cells at room temperature at 5300 rpm for 10 minutes. The supernatant was discarded, and the cells pellets snap frozen in liquid nitrogen. Cell pellets were then stored at -80 °C until RNA was extracted.

**4.2.2 RNA extraction**

RNA extraction from the cultures was performed by Vaibhav Bhardwaj and Katherine Helliwell using Bioline RNA Easy Kit (http://www.bioline.com/uk/isolate-ii-rna-mini-kit.html). The manufacturer’s protocol was followed with additional steps of pre-treating cells with lysozyme to break down bacteria cell walls. RNA quality was assessed using a nano-drop and bioanalyser.

**4.2.3 RNA preparation and Sequencing**

Preparation and sequencing of the RNA samples was performed by Cambridge Genomic Services, in the Department of Pathology, University of Cambridge by Dr Edward Farnell. The samples were prepared using the InDA-Cr RNA removal kit from NuGen (http://www.nugen.com/products/rna-seq), which serves to deplete rRNA sequences. The kit was custom-made to remove rRNA from *M. loti* and *C. reinhardtii* (as no sequences were available for *L. rostrata* rRNA). Sequencing of the samples was performed using Illumina NextSeq on High output mode. Sequencing was set to generate 75 bp pair-end reads.

**4.2.4 Read filtering and QC**

The raw reads from all 12 samples (Table 4.1) were subjected to filtering steps to establish a high-quality dataset for assembly.

The four co-culture samples were subjected to a filtering step, where all reads from each sample were aligned to the *M. loti* transcriptome using Bowtie2 on the ‘sensitive’ setting. This enabled separation of reads from *M. loti* and *L. rostrata*. The *M. loti* transcriptome was obtained from the Ensembl database (Kopylova et al, 2012).

The rRNA derived reads were then removed using SortMeRNA (Kopylova et al, 2012). Using this software all reads from the samples were aligned to the
databases SiLVA (Quast et al, 2013) and rfn (Nawrocki et al, 2015), which contain both eukaryotic and prokaryotic rRNA sequences. SortMeRNA uses the BLAST algorithm to find and align reads to the rRNA databases and matching reads (using default e-value cutoff of 1) were removed from the dataset used for subsequent analysis.

Next, low quality reads, unpaired reads and Illumina adapter sequences were removed using Trimmomatic (Bolger et al, 2014). Trimmomatic was set up to remove reads with low quality in the first two and last two bases, as well as reads below 25 bp. These settings were based on recommendations by MacManes (2015). Using Trimmomatic, unpaired reads were identified and removed from the dataset.

The quality of the final set of reads was assessed using FastQC with default settings (Andrews, 2016).

**4.2.5 De novo transcriptome assembly for L. rostrata**

The assembly process was performed based on advice from Chris Boursnell, Department of Plant Science, University of Cambridge.

The de novo assembly was based on filtered reads from both the axenic *L. rostrata* cultures and the co-cultures. Three different assemblies were created using the assemblers, IDBA-tran (Peng et al, 2013), Trinity (Haas et al, 2013) and SOAPdenovo-Trans (Xie et al, 2014). IDBA was used with the settings: minimum K-mer length 21, max K-mer length 69 and a 0.98 similarity cut-off. Trinity was run with default settings. SOAPdenovo-Trans was used with default settings and a K-mer length of 3. The quality of each assembly was analysed using Transrate (Smith-Unna et al, 2016) using default settings. A merged assembly was created using Transfuse (Boursnell, 2016) using default settings. The final assembly was based on Transrate to identify the subset of contigs from the three different assemblies that gave the best Transrate score.

**4.2.6 Quantification of contigs expression and identification of differentially expressed contigs**

Reads from the axenic *L. rostrata* samples and the co-culture samples were mapped to the final assembly using the short read mapper Bowtie 2 (Langmead & Salzberg, 2012). Bowtie 2 was run on ‘sensitive’ mode. The mapped reads
were quantified to get the number of reads mapping to each of the contigs in the final assembly. This was done using RSEM (Li & Dewey, 2011) with default settings. Statistical analysis of the read counts was done using DeSeq2 (Love et al., 2014) using default settings. A P-value adjusted for multiple testing was used to identify differentially expressed contigs and a cutoff of $P_{\text{adjusted}} < 0.05$ was used.

**4.2.7 Annotation of the de novo assembly**

Annotation of the assembly was done based on homology between the constructed contigs and annotated sequences from public databases. The sequence aligner Diamond (Buchfink et al., 2015) was used to identify the closest protein homolog in NCBI NR (NCBI Resource Coordinators, 2016) for each contig. Diamond uses an optimised version of BLASTX (Altschul et al., 1997) to identify similarities between contigs and protein sequences. Diamond was run using an e-value threshold of $< 10^{-5}$ and “sensitive” setting. The contigs were also compared to the annotated transcriptome of *C. reinhardtii* using CRB-BLAST (Aubry et al., 2014) and the same settings as described above. The *C. reinhardtii* transcriptome sequences and annotation was obtained from Ensembl Plants (Bolser et al., 2017). In addition, translated versions of the contigs were analysed for the presence of conserved proteins domains. Blast2GO (Götz et al., 2008) was used to run InterProScan (Mitchell et al., 2015), which searched for domains in the databases Interpro, Pfam and TIGRFAM (Haft et al., 2013). InterProScan was run with default settings. Based on the results from Diamond/BlastX, GO-terms (Ashburner et al., 2000) were added to each contig where possible. This was done using Blast2GO and the mappings provided by The Gene Ontology Consortium (2015). Based on the identified homologs for each contig, the corresponding organism for each homolog was found. This was done by combining the results from Diamond/BLASTX results and mappings from NCBI Taxonomy (NCBI Resource Coordinators, 2016).

**4.1 Results**

In the sections below, I will both describe the results from the RNA extraction, sequencing process, *de novo* assembly process, annotation process and the biological findings.
4.1.1 RNA quality from extraction

RNA was extracted from four replicates of axenic *L. rostrata* cultures, axenic *M. loti* cultures and co-cultures of the two organisms. Table 4.1 shows the total RNA extracted from each culture as well as the ratios of A260/A280, A260/A230 and RNA integrity number (RIN). The A260/A280 ratio is used to indicate contamination from proteins, while the A260/A230 ratio is used to control for polysaccharide contamination (Asif *et al.*, 2000). Generally, samples should have values of 1.8 or higher in order to be considered contamination free (Asif *et al.*, 2000; Baelde *et al.*, 2001). *L. rostrata* and co-culture samples were well above this recommended threshold, but the *M. loti* samples were only slightly above the threshold. Another measurement used to quantify the samples quality is the RIN. The RIN is used to identify degraded RNA samples and it is calculated based on microcapillary electrophoretic RNA separation and an algorithmic approach to do so (Schroeder *et al.*, 2006). The RIN is given as a number between 1 (completely degraded) and 10 (completely intact) and typically a minimum of 8 is required for RNA-seq protocols (Van Verk *et al.*, 2013). Both the *M. loti* and *L. rostrata* samples were above this threshold, but the co-cultures samples appeared to have been partly degraded.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Total RNA [ng]</th>
<th>A260/A280</th>
<th>A260/A230</th>
<th>RIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-culture #1</td>
<td>3991.2</td>
<td>2.19</td>
<td>2.16</td>
<td>7</td>
</tr>
<tr>
<td>Co-culture #2</td>
<td>3704.4</td>
<td>2.19</td>
<td>2.33</td>
<td>5.9</td>
</tr>
<tr>
<td>Co-culture #3</td>
<td>3934.8</td>
<td>2.18</td>
<td>2.27</td>
<td>6.2</td>
</tr>
<tr>
<td>Co-culture #4</td>
<td>4186.8</td>
<td>2.19</td>
<td>2.4</td>
<td>5.3</td>
</tr>
<tr>
<td><em>L. ros.</em> axenic #1</td>
<td>6622.8</td>
<td>2.23</td>
<td>2.22</td>
<td>8.2</td>
</tr>
<tr>
<td><em>L. ros.</em> axenic #2</td>
<td>4075.2</td>
<td>2.19</td>
<td>2.36</td>
<td>8.7</td>
</tr>
<tr>
<td><em>L. ros.</em> axenic #3</td>
<td>6795.6</td>
<td>2.21</td>
<td>2.15</td>
<td>8.7</td>
</tr>
<tr>
<td><em>L. ros.</em> axenic #4</td>
<td>2919.6</td>
<td>2.21</td>
<td>2.45</td>
<td>8.3</td>
</tr>
<tr>
<td><em>M. loti</em> axenic #1</td>
<td>1534.8</td>
<td>1.85</td>
<td>2.18</td>
<td>8.4</td>
</tr>
<tr>
<td><em>M. loti</em> axenic #2</td>
<td>942</td>
<td>1.81</td>
<td>2.27</td>
<td>9.7</td>
</tr>
<tr>
<td><em>M. loti</em> axenic #3</td>
<td>916.8</td>
<td>1.84</td>
<td>2.03</td>
<td>10</td>
</tr>
<tr>
<td><em>M. loti</em> axenic #4</td>
<td>1030.8</td>
<td>1.84</td>
<td>2.05</td>
<td>9.5</td>
</tr>
</tbody>
</table>
4.1.2 Quality of raw sequence data

Notwithstanding the quality measures of the RNA samples, all 12 were sequenced using Illumina NextSeq with 75 bp read length at the Cambridge Genomic Services, Dept. Pathology, University of Cambridge. The number of reads returned from each sample is listed in Table 4.2, and on average each sequence dataset contained 22.7 million reads and 1.70 billion nucleotides.

For each sample, a quality report of the sequence output was created using FASTQC (Andrews, 2016). While the reports contain multiple outputs, only the quality scores of the base calling is shown in Figure 4.1 and for just one sample from each of the three experimental setups. The full reports are available on request due to their large size. The quality graphs from the other samples are similar to the ones shown here. In collaboration with Dr Krys Kelly, Dept. of Plant Science, University of Cambridge, the reports from FASTQC were interpreted, and the sequencing data was found to be without any issues and of sufficient quality.

Table 4.2– Number of raw reads from each sample from the RNA sequencing step

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Number of raw reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-culture #1</td>
<td>19,115,639</td>
</tr>
<tr>
<td>Co-culture #2</td>
<td>23,984,157</td>
</tr>
<tr>
<td>Co-culture #3</td>
<td>25,252,132</td>
</tr>
<tr>
<td>Co-culture #4</td>
<td>23,893,172</td>
</tr>
<tr>
<td>L. ros. axenic #1</td>
<td>22,486,581</td>
</tr>
<tr>
<td>L. ros. axenic #2</td>
<td>18,826,928</td>
</tr>
<tr>
<td>L. ros. axenic #3</td>
<td>20,730,016</td>
</tr>
<tr>
<td>L. ros. axenic #4</td>
<td>19,671,265</td>
</tr>
<tr>
<td>M. loti axenic #1</td>
<td>22,687,148</td>
</tr>
<tr>
<td>M. loti axenic #2</td>
<td>25,059,376</td>
</tr>
<tr>
<td>M. loti axenic #3</td>
<td>26,746,657</td>
</tr>
<tr>
<td>M. loti axenic #4</td>
<td>24,458,223</td>
</tr>
</tbody>
</table>
Figure 4.1 – Average quality score for base calling on each position across the 75 bp length reads from the RNA-seq dataset. The quality scores were calculated using FASTQC (Andrews, 2016)
4.1.3 Read filtering and sorting

4.1.3.1 Sorting reads from co-cultures

In order to process the reads from co-cultures, reads from \textit{M. loti} were separated from the rest of the reads. This was done by aligning the raw sequence reads from the four co-culture samples to a reference transcriptome for \textit{M. loti} with the results shown in Table 4.3. The sorting revealed that just 0.06% of the reads from the co-culture samples could be aligned to the \textit{M. loti} transcriptome. This small number of reads did not allow further analysis of the \textit{M. loti} transcriptome and the rest of the study is therefore focused on the \textit{L. rostrata} transcriptome.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total reads</th>
<th>Reads from \textit{M. loti}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-culture #1</td>
<td>19,115,639</td>
<td>13,189</td>
</tr>
<tr>
<td>Co-culture #2</td>
<td>23,984,157</td>
<td>12,167</td>
</tr>
<tr>
<td>Co-culture #3</td>
<td>25,252,132</td>
<td>15,819</td>
</tr>
<tr>
<td>Co-culture #4</td>
<td>23,893,172</td>
<td>14,135</td>
</tr>
</tbody>
</table>

Table 4.3 – Total number of reads from each co-culture sample and the number of reads aligning to the \textit{M. loti} transcriptome from those samples. The average \textit{M. loti} fraction of reads is 0.06% across the samples.

4.1.3.2 rRNA depletion

As up to 90% of RNA in a cell is rRNA, separating rRNA from mRNAs is essential when attempting to study gene expression (Kopylova \textit{et al}, 2012). As mentioned in the methods section, a custom rRNA removal kit was applied prior to sequencing. While this improves the rRNA to mRNA ratio, a computational filtering of the sequenced reads can reduce the rRNA level further. The computational tool SortMeRNA (Kopylova \textit{et al}, 2012) was used to compare the sequenced reads to rRNA sequences from SILVA (Quast \textit{et al}, 2013) and rFam (Nawrocki \textit{et al}, 2015) and remove rRNA reads (Results in Table 4.4). The results showed that each sample contained 80.3% rRNA reads on average, which is comparable to other studies. For example, in a study of the brown alga \textit{Ectocarpus siliculosus}, 94% of all reads from the RNA-seq were found to be rRNA (Saint-Marcoux \textit{et al}, 2015), whilst for a newly discovered \textit{Chlorella} species, the rRNA fraction was found to be between 60-89% of the sequenced reads (Mansfeldt \textit{et al}, 2016).
Table 4.4 – Reads from each sample after *in silico* removal of rRNA and Trimmomatic filtering. The removal of rRNA sequences was after aligning reads to databases of rRNA sequences. The Trimmomatic filtering step removed low quality reads and adapter sequences.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Initial reads</th>
<th>After rRNA removal</th>
<th>After Trimmomatic filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. ros. axenic #1</td>
<td>22,486,581</td>
<td>4,092,435</td>
<td>3,718,777</td>
</tr>
<tr>
<td>L. ros. axenic #2</td>
<td>18,826,928</td>
<td>2,791,225</td>
<td>2,553,228</td>
</tr>
<tr>
<td>L. ros. axenic #3</td>
<td>20,730,016</td>
<td>3,636,041</td>
<td>3,268,737</td>
</tr>
<tr>
<td>L. ros. axenic #4</td>
<td>19,671,265</td>
<td>2,793,438</td>
<td>2,541,483</td>
</tr>
<tr>
<td>Co-culture #1</td>
<td>19,102,450</td>
<td>4,419,068</td>
<td>3,984,912</td>
</tr>
<tr>
<td>Co-culture #2</td>
<td>23,971,990</td>
<td>5,724,279</td>
<td>5,246,087</td>
</tr>
<tr>
<td>Co-culture #3</td>
<td>25,236,313</td>
<td>5,879,262</td>
<td>5,377,162</td>
</tr>
<tr>
<td>Co-culture #4</td>
<td>23,879,037</td>
<td>5,463,412</td>
<td>4,994,873</td>
</tr>
</tbody>
</table>

4.1.4 Quality filtering and trimming of adaptors

While the quality of the base calling in the reads from the sequencing process in this study was high, it is still recommended to remove reads where the base calling is of low quality, the reason being that errors in base calling can affect the mapping and assembly processes (Conesa *et al*, 2016). Adaptor sequences from the experimental steps of the process can also cause unwanted disturbances and should therefore also be removed. Using the software Trimmomatic (Bolger *et al*, 2014), low quality reads and adaptors were removed from pools of reads from all samples. The result of this process is shown in Table 4.4. On average 9.9% of the reads from each sample were removed in this filtering step, which is on par with other recent RNA-seq studies. In a transcriptomics study of the alga *Euglena gracilis*, this filtering step removed 8.3% of the raw reads (Yoshida *et al*, 2016), whilst 16.0% of raw reads were removed in a study of the fruit *Garcinia mangostana* (Matra *et al*, 2016).

4.1.5 De novo assembly of *L. rostrata* transcriptome

As no transcriptome or genome exists for *L. rostrata*, one of the first objectives was to build an assembly representing the transcriptome from the cleaned sequence reads. To achieve the best possible assembly, three different tools for *de novo* transcriptome assembly were utilised and the resulting assemblies were combined. Using the assemblers, IDBA-tran (Peng *et al*, 2013), Trinity (Haas *et al*, 2013) and SOAPdenovo-Trans (Xie *et al*, 2014), three different assemblies were created, each with a different number of contigs (Table 4.5). Using the software Transrate (Smith-Unna *et al*, 2016), the quality of each contig in each assembly was assessed based on mapping the reads to back to the contigs. For each assembly, an aggregated score based on the quality of the contigs was also calculated (Table 4.5). The tool Transfuse (Boursnell, 2016) was used to merge
the three assemblies, which resulted in a Transrate score of 0.540. In order to provide context for the scores, Smith-Unna et al (2016) analysed 155 published transcriptomes and found that 50% of the studies had a Transrate score above 0.22. The final step in the filtering process was removal of low-expressed contigs, as performing a differential expression analysis using these will not produce high confidence results (Vijay et al, 2013). First, expression of each contig was quantified by mapping the cleaned reads to the contigs using Bowtie2 (Langmead & Salzberg, 2012) and quantification of read counts was done using RSEM (Li & Dewey, 2011). The transcriptome assembly was then filtered using threshold of >1 TPM for each transcript. The final assembly had 134,682 contigs, a Transrate score of 0.548 and a length distribution as shown in Figure 4.2a.

Table 4.5 – Size and quality of the three assemblies made, the merged assembly and the final assembly. The final assembly is based on a subset of the merged assembly, selected based on Transrate scores.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>No. of contigs</th>
<th>Transrate scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDBA-tran</td>
<td>148,995</td>
<td>0.371</td>
</tr>
<tr>
<td>Trinity</td>
<td>175,234</td>
<td>0.178</td>
</tr>
<tr>
<td>SOAPdenovo-Trans</td>
<td>196,871</td>
<td>0.221</td>
</tr>
<tr>
<td>Merged assembly</td>
<td>199,550</td>
<td>0.540</td>
</tr>
<tr>
<td>Final assembly</td>
<td>134,682</td>
<td>0.548</td>
</tr>
</tbody>
</table>
To provide context for the distribution of contig sizes, I extracted the transcriptome for *V. carteri* and *C. reinhardtii* from ENSEMBL (Bolser et al., 2017) and plotted the distribution in Figure 4.2b. When comparing the distributions, it was found that 59.3% of the *L. rostrata* contigs were <500 nt and 7.6% were >2000 nt, compared to 12.4% and 43.5% for *C. reinhardtii* transcripts. This could mean that some of the transcripts from the *L. rostrata* cultures have been split into multiple contigs. This seems even more likely considering that the *L. rostrata* assembly has ~10x more contigs than the number of *C. reinhardtii* and *V. carteri*
transcripts. While this is suboptimal, it is also to be expected when building a de novo assembly and it is similar to what has been found in other studies of algae without a sequenced genome (O’Neill et al., 2015; Yoshida et al., 2016).

### 4.1.6 Annotation

Annotation of the 134,682 contigs in the final assembly was carried out using sequence similarity between the contigs and annotated sequences from public databases. The software tool DIAMOND (Buchfink et al., 2015) enables the BLASTX search algorithm to be applied on larger datasets such as this by speeding up the process >1000x while maintaining a similar sensitivity. This allows a complete homology search to be completed in hours rather than months. For 35,633 contigs (~26%) at least one match was found in the NCBI NR database. Using BLAST2GO (Götz et al., 2008) Gene Ontology (GO) terms were assigned to the contigs where possible. Based on the identified homologs from NCBI NR, KEGG terms (Kanehisa et al., 2014) were assigned to contigs using MEGAN (Huson et al., 2016). Using CRB-BLAST, the closest reciprocal matches between each contig and the *C. reinhardtii* transcriptome were identified. The number of contigs annotated using each method as well as examples of annotation is shown in Table 4.6. Two important caveats to note are that 1) each contig can potentially have two different annotations due to the differences between the methods and 2) multiple contigs can have the same annotation. Splice variants and false variants caused by the assembly process are some of the reasons for multiple contigs ending up with the same annotation. Both issues are to be expected in de novo assembly projects (Ono et al., 2015).
Table 4.6 – Overview of annotation results. Annotations of contigs were made using DIAMOND to identify related sequences from NCBI NR. From the NCBI NR results, GO and KEGG categories were assigned. CRB-BLAST was used to identify homologs from the C. reinhardtii transcriptome. Examples are provided in the second part of the table.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of contigs annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>16,206</td>
</tr>
<tr>
<td>KEGG</td>
<td>7,998</td>
</tr>
<tr>
<td>C. reinhardtii transcriptome</td>
<td>15,603</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contig</th>
<th>KEGG Annotation</th>
<th>GO annotation</th>
<th>C. reinhardtii homolog</th>
</tr>
</thead>
<tbody>
<tr>
<td>contig15697.1</td>
<td>K00012 UDPglucose 6-dehydrogenase [EC:1.1.1.22]</td>
<td>UDP-glucose 6-dehydrogenase, eukaryotic type  tr</td>
<td>A8J914</td>
</tr>
<tr>
<td>contig3789.1</td>
<td>K01581 ornithine decarboxylase [EC:4.1.1.17]</td>
<td>Ornithine decarboxylase        tr</td>
<td>A8J7E8</td>
</tr>
<tr>
<td>contig21177.1</td>
<td>K01738 cysteine synthase A [EC:2.5.1.47]</td>
<td>Cysteine synthase CysK           tr</td>
<td>A8ISA9</td>
</tr>
</tbody>
</table>

GO-terms can be used to give an overview of the types of contigs identified in the dataset. 16,206 contigs had at least one GO term associated with them and across all contigs 103,082 GO terms were identified. The identified GO-terms were aggregated into groups based on the GO taxonomy and accumulated counts for each group can be found in Figure 4.3. The figure provides an overview of the assembly and shows that a broad range of predicted functions can be found in the constructed assembly. GO-term mapping show that a large number of contigs are classified as being involved catalytic and metabolic processes. A smaller set of contigs were mapped to process related to cell reproduction, transport, transcription regulation and cell death. These observations are to be expected from a good representation of a transcriptome and increases confidence in the assembly as a truthful representative of the actual L. rostrata transcriptome.
Figure 4.3 - Accumulated counts for each GO-term. Level 2 of the GO taxonomy was used and terms at lower levels were summed to their Level 2 term. The counts are based on GO annotation of 16,206 contigs with a total of 103,082 GO terms.

### 4.1.7 Comparison with other organisms

There is no public genome of *L. rostrata* and this study is the first time the *L. rostrata* transcriptome has been sequenced. This is therefore also the first study of how *L. rostrata* is phylogenetically related to other sequenced or partly sequenced organisms. The goal of this analysis was to identify species that are closely related to *L. rostrata* using the constructed assembly for *L. rostrata*. The final de novo assembly of 134,682 contigs was used and for each of the contigs, the closest matching sequence from NCBI NR was identified using BLASTX. The organism from which the matching sequence originated was identified using NCBI Taxonomy. This was done for each contig and summed up across the whole assembly. To reduce the noise created by errors in the assembly process, the results were weighted using the read count for each contig. The results show that the 85% of the mapped reads map to contigs that have a close homolog in either *C. reinhardtii*, *C. incerta* or *V. carteri*, which are all part of the order *Chlamydomonadales* (*Volvocales*) (Figure 4.4). While not surprising, this firmly places *L. rostrata* among the green algae, but as *V. carteri* and *C. reinhardtii* stem from two different families among the *Chlamydomonadales* the data do not enable a more specific classification. Surprisingly, 2.29% of the reads comes from contigs where the closest sequence match is from *Klebsormidium flaccidum* from the *Streptophyta* phylum. The contigs that have the closest match in *K. flaccidum* were inspected in closer detail to identify the types of genes that *L.*...
rostrata appears to share with *K. flaccidum*. However, based on the KEGG and GO annotation of the contigs, no apparent pattern could be found.

Among the top ten species from the analysis are eight algae, but also two non-algal species. *Sandaracinus amylyticus* is a delta-proteobacteria from the *Myxococcales* order and 0.61% of the reads from the experiment map to contigs with a high similarity to transcripts from this bacterium. *Octopus bimaculoides* is also known as a California two-spot octopus and 0.54% of the reads map to contigs related to this organism. In the case of both of these organisms no further investigation of these contigs was performed in this study.

![Figure 4.4 – Fraction of the reads assigned to each species. Each read was assigned to a species with a public genome based on homology using BLASTX. The figure shows that 43.38% of the reads from the experiment map to sequences with a close homolog in *C. reinhardtii*.](image)

### 4.1.8 Differentially expressed contigs

Using the number of reads mapping to each contig and the software DeSeq2 (Love *et al*, 2014), 577 contigs were identified as having a statistically significant (\( p_{adj} < 0.05 \)) differential expression between the axenic *L. rostrata* and co-culture samples. Of the 577 contigs, 368 were expressed more (up-regulated) in co-cultures compared to the axenic cultures, while 209 were expressed at lower levels (down-regulated) in the axenic cultures compared with co-cultures. For the
down-regulated contigs it was possible to assign GO-terms to 43 of them and for the up-regulated contigs GO-terms were assigned to 130. A summary of the GO-terms assigned to the differentially expressed contigs can be found in Figure 5. It should be noted that for most categories such as “nucleotide binding”, “transferase activity” and “nitrogen compound metabolic process” there are several contigs that are up-regulated and several contigs that are down-regulated.
4.1.8.1 Increased expression of transporter for plant hormone indole-3-acetic acid

Analysis of the different classes of differentially expressed genes and the significance for the behaviour of *L. rostrata* in co-culture is ongoing in our laboratory. I focused on two differentially expressed genes that can be related to other studies in the literature. Contig7872.4 appears to be expressed twice as
high in co-cultures ($P_{\text{adjusted}} = 0.01$) as in axenic *L. rostrata* cultures. The homology search carried out to identify the function of the related enzyme showed that the contig has a very high similarity to an MDR-like ABC transporter from *C. reinhardtii* and the ABCB4 transporter from *A. thaliana*. The ABCB4 transporter has been showed to be a transporter for uptake of the plant hormone indole-3-acetic acid (IAA), which is a type of auxin (Santelia *et al.*, 2005; Terasaka *et al.*, 2005). IAA is known to regulate growth in plants and has been shown to stimulate growth of the marine algae *P. multiseries* (Amin *et al.*, 2015). The hormone has been shown to be produced by multiple alpha-proteobacteria and in particular *rhizobia* have been associated with synthesis of this molecule (Theunis *et al.*, 2004; Jones *et al.*, 2007), although IAA has not been found to have any metabolic function in bacteria (Lau *et al.*, 2009). Based on a lookup in KEGG, *M. loti* appears to have the required genes encoding the enzymes for IAA synthesis (*mlr0712, mll3668* and *mll2219*) using the tryptophan dependent IPA pathway (Kanehisa *et al.*, 2014). Further to this, there is experimental evidence that show that the related species *Mesorhizobium ciceri*, *Sinorhizobium meliloti* and *Rhizobium sp. NGR234* can produce IAA (Ahmad *et al.*, 2008; Duca *et al.*, 2014). In a study of a co-culture of the diatom *P. multiseries* and the alpha-proteobacterium *Sulfitobacter SA11*, it was found that the bacterium increased expression of IAA synthesis related genes by 2 to 4-fold when it was co-cultured with the diatom.

### 4.1.8.2 Increased expression of the kinase SNRK2.1 in co-cultures

Contig7908.1 has a 15-fold increase in expression in co-cultures compared with mono-cultures ($p_{\text{adj}} = 5.3*10^{-8}$). Based on the homology analysis, the contig appears to represent the gene encoding a SNRK2.1 enzyme. SNRK2 is a family of serine/threonine kinases related to abiotic stress and in algae SNRK2.1 in particular has been associated with sulphur deprivation (Gonzalez-Ballester *et al.*, 2008; Gonzalez-Ballester *et al.*, 2010; Kulik *et al.*, 2011). Under sulphur replete conditions SNRK2.1 is inhibited by SNRK2.2, but under sulphur deplete conditions SNRK2.1 activates genes that enable the cell to adapt to the new condition (Gonzalez-Ballester *et al.*, 2010). In a set of transcriptomics experiments with a *C. reinhardtii* ΔSNRK2.1 mutant it was shown that under sulphur deplete conditions cell viability was greatly reduced (Gonzalez-Ballester *et al.*, 2010).
From the transcript data it was found that absence of the SNRK2.1 reduces the expression of methionine synthase (METH), biotin synthesis transcripts (BIOB1 and BIOB2) and thiamine synthesis transcripts (THI4 and THIC) in sulphur deplete conditions. Other aspects of metabolism that appear to be affected by SNRK2.1 include photosynthesis and amino acid synthesis. It is not currently understood what physiological effects an increase in SRNK2.1 expression will cause in an alga.

4.2 Discussion

4.2.1 Transcriptome analysis of a culture of algae and bacteria
From the initial cultures of L. rostrata, it was possible to conduct a set of experiments and data processing steps that have led to the first representative transcriptome for this alga. In the experimental steps of the process, suitable protocols for RNA-extraction were identified and developed as needed by collaborators Katherine Helliwell and Vaibhav Bhardwaj. When trying to compare to cultures grown under two different conditions (axenic vs. co-culture) it was critical to evaluate and select conditions that allowed comparison of the two samples. The experimental conditions under which the cultures were grown could have a significant impact on the amount of RNA extracted. As growth rates differ under the two conditions, it was important to select time points where the cultures were in a similar state (exponential, steady state, or decline).

In the quality control of the RNA extracted from the 12 cultures, the co-culture samples appeared to have some level of degradation based on the RIN (between 5.3 and 7). A decision was made to proceed with these samples despite these values being below ideal level for an RNA-seq experiment. The samples in question contained chloroplast, mitochondrial, eukaryotic and prokaryotic RNA and it is possible that influences the calculation of the RINs. To my knowledge, no other study of co-cultures of an alga and a bacterium have reported RINs, and it is still uncertain if this is a good measurement for co-cultures.

Post sequencing, the reads from co-cultures were aligned to the M. loti transcriptome to identify the fraction of M. loti transcripts. Disappointingly, the M. loti transcripts made up less than 0.1% of the total reads and there were not enough reads to do a transcriptome analysis of M. loti from the co-culture
samples. Although data were available for the *M. loti* transcriptome from the axenic *M. loti* cultures, analysis was postponed as there was no other conditions to compare it with. In this experiment, we had aimed to get sequence for both organisms from one set of co-culture samples. For future efforts in sequencing algae and bacteria co-cultures, it would be advisable to carefully investigate the ratio of algae to bacteria RNA in the samples prior to sequencing. This could be done by calculating the RNA per cell for both the algae and the bacteria and then correlate that with the cell counts for the co-cultures to calculate the ratio of total RNA. In cases where it is deemed unfeasible to get sequencing material for both organisms from a single sample, one possibility would be to extract eukaryotic and bacterial mRNA from separate samples of the co-culture and sequence them individually. Eukaryotic mRNA could be extracted using a simply Poly-A kit, to capture RNAs with a polyadenylated tail. For extraction of bacterial mRNA, the process will likely include removal of poly-A sequences, prokaryotic rRNA and eukaryotic rRNA (Kumar *et al*, 2016). Another approach to solving this issue could be to use a filtering step of the co-cultures to separate bacteria from algae. This has successfully been done using a 5 µm filter (Eigemann *et al*, 2013). In the proteomics study, that I co-authored, where the same model system was studied, we encountered a similar problem: it was not possible to identify proteins originating from *M. loti* from the co-culture samples. The lack of reads mapping to *M. loti* is therefore unlikely to be due to errors in the approach used, but rather due to a very high *L. rostrata:* *M. loti* RNA ratio.

Creating a *de novo* assembly of a transcriptome from individual reads is not a trivial task, and care should go into the choice of method. Currently, there is not a single method that can claim to be superior, and multiple methods of assembly, and permutations to these, exist. A central challenge in creating a good representation of the transcriptome from short reads is that it is not possible to determine exactly how accurately the assembly matches the actual transcriptome.

The 134,682 contigs in the final assembly from *L. rostrata* is ~10x higher than the number of transcripts in the *C. reinhardtii* transcriptome. While this can appear surprising, it is on par with other comparable studies. In a transcriptomics study of the green algae *Euglena gracilis*, the *de novo* assembly contained 113,295
contigs (Yoshida et al, 2016) and in another study of E. gracilis published only a few months prior the assembly was made up of 233,748 contigs. To reduce the number of contigs derived from the assembly process to a level that resembles the actual transcriptome, researchers sometimes resort to filtering of the contigs. In a study of the alga Dunaliella salina, an assembly with 43,236 contigs was created, but this number was only achieved after removing contigs shorter than 300 base pairs and contigs with fewer than 40 reads per contig (Fang et al, 2017). Similarly, in a study of the green alga Tetraselmis sp. KCTC12432BP the initial assembly of 425,485 contigs was reduced to 26,245 contigs by filtering out those without a homolog in NR protein database (Shin et al, 2016). While these filtering tactics are practical and effective at reducing the number of contigs, the thresholds used to do the filtering are often arbitrary and rely on intuition rather than any reproducible method. Methods using a formal approach to reduce the number of contigs from de novo assembly have been developed but are still novel. De novo RNA-Seq Assembly Pipeline (DRAP) is a data pipeline project that aims to increase the compactness and correctness of assemblies from Trinity and Oases (Cabau et al, 2017). The assemblies are compacted by extracting the longest and most covered contig from sets of highly similar contigs, using CD-HIT-EST to merge similar contigs, and Transrate (Smith-Unna et al, 2016) to remove contigs deemed to be of low quality. While this processing pipeline also relies on thresholds, the authors have demonstrated that they are able to reduce the number of contigs in several assemblies without reducing the completeness of the assembly.

For annotation of de novo assemblies again there is no single method or pipeline that have proven to perform best in all cases. Due to the number of contigs in assemblies typically being in the tens of thousands or greater, manual comparisons between the sequences from the assembly and sequences from some annotated source is not feasible. Instead most researchers rely on an automated approach to generate annotations. Currently, most research projects use some version of the BLAST algorithm to identify sequences with a high degree of similarity to the contigs in question and then transfer the annotations. The source of the annotated sequences varies, but are generally combinations of NCBI NR, UniProt/SwissProt and transcriptomes from model organisms such
as *A. thaliana*. Domain identification tools such as InterProScan (Jones *et al*, 2014) are used to search protein domain databases such as Pfam. In addition to these, KEGG and GO annotations are often added on the basis of the results from BLAST based annotation using the mapping system provided by their host organisations. While semi-automated, the methods above still require the user to be familiar with a command-line Unix environment, so software packages to automate this process have been developed. Tools such as Blast2GO (Götz *et al*, 2008) gives users a graphical interface that allows them to carry out some forms of annotation. Other tools that have aimed to simplify the annotation process are AfterParty (Jones & Blaxter, 2013), Annot8r (Schmid & Blaxter, 2008) and Trinotate ([https://trinotate.github.io/](https://trinotate.github.io/)). In this study, 24,102 total contigs from the *L. rostrata* assembly were given some level of annotation. Compared to the number in the assembly (134,682 contigs), this might appear underwhelming. However, while it might be possible to apply more tools to generate more annotations, the current status is on par with recent and similar studies such as the previously mentioned study of *E. gracilis*, where 12,020 out of 233,748 contigs were annotated with a GO-term (O’Neill *et al*, 2015).

### 4.2.2 Differentially expressed contigs and further experiments

As presented in the results section, a set of contigs with differential expression was identified from the transcriptomics experiment. In the results section I presented two examples of contigs that are potentially contributing to mutualistic interaction observed between *L. rostrata* and *M. loti*

To my knowledge, this study is the first time an IAA transporter in an alga has been observed to be expressed at higher levels in the presence of a bacterium. IAA has long been known to be central to plant development, where it acts as a hormone and is involved in central processes such as shoot elongation and root development (Leyser, 2010). IAA is known to be produced by several types of soil bacteria known to be associated with plant roots such as species of *Pseudomonas, Rhizobium, Azospirillum, Enterobacter, Azotobacter, Klebsiella, Alcaligenes, Pantoea* and *Streptomyces*. IAA synthesis is also found in plants and fungi (Duca *et al*, 2014). The next step from this point would be to set up a set of experiments to verify and this observation. One way of verifying the
observation is to set up a set of axenic- and co-cultures of *L. rostrata* and *M. loti* and then carry out a qPCR experiment targeting the putative IAA transporter to verify that expression is indeed increase when the bacteria is present. It would also to be central to verify the function of enzyme corresponding to Contig7872.4. *L. rostrata* is currently not suited for genetic modifications, such as expression of transgenes or knockout experiments. To address this, *C. reinhardtii* potentially could be used as the *C. reinhardtii* gene CHLREDRAFT_138672 is highly similar to Contig7872.4 (75% identity based on a nucleotide alignment) and thus likely to be an IAA transporter. While Amin *et al* (2015) demonstrated that IAA could enhance the growth rate of *P. multiseris*, the effects of IAA and auxins in general are still poorly understood in algae. Amin *et al* (2015) hypothesized that IAA is one among many molecules that are used by algae and bacteria to reinforce mutualistic behaviour. Based on the assumption that the IAA transporter is in fact up-regulated in co-cultures and that IAA is central to the observed mutualism, it would be interesting to study the effect on co-cultures using *C. reinhardtii ΔmetE* mutant (Helliwell *et al*, 2014a). Using this strain with the IAA transporter being knocked down using miRNA (Molnar *et al*, 2009) could be used to test if IAA uptake is central to the stability of the mutualistic interaction.

The other highlighted contig with up-regulated expression in co-cultures is Contig7908.1, which is a putative kinase called SNRK2.1. This kinase is thought to influence the expression of a wide array of genes in algae and experiments have been undertaken to demonstrate the effect of absence of SNRK2.1 (González-Ballester *et al*, 2010). The next step forward from this initial observation is to confirm that Contig7908.1 corresponds to SNRK2.1 in *L. rostrata* and that the gene is indeed expressed at different levels between co-cultures and axenic cultures. Following on from that it would interesting to investigate if *L. rostrata* or an equivalent alga with a ΔSNRK2.1 mutation would be able to form a mutualistic interaction with *M. loti*.

Further analysis that would provide validation and context for this study is a detailed comparison of the data from the mentioned proteomics study of *M. loti* and *L. rostrata* (Helliwell *et al*, 2017, accepted for publication). In this study 588 algal proteins were identified with 153 of them being differentially expressed. From this it was identified that two of the main areas of metabolism in the algae
that appear to be affected in co-cultures are photosynthesis and amino acid metabolism. Interestingly, these areas have been shown to change expression of transcripts in a *C. reinhardtii ΔSNRK2.1* strain (González-Ballester *et al*, 2010). An exhaustive understanding of the mechanisms underpinning the mutualistic interaction between *L. rostrata* and *M. loti* has yet to be achieved. However, through the experiments conducted in this study I have created a new resource for identifying candidate genes central to the co-cultures between alga and bacteria. From the list of differentially expressed contigs, two in particular were identified and further experiments should be undertaken to confirm the hypotheses generated from this study. Further investigation of the differentially expressed contigs will reveal more clues as to which genes are underpinning the observed interactions between algae and bacteria.
5 Discussion and conclusion

5.1 Overview

Algae are responsible for as much as half of the primary production on the planet (Whitman et al., 1998; Field et al., 1998; Amin et al., 2015). In oceans, algae are surrounded by a phycosphere containing secreted materials as well as associated bacteria, and the interaction between algae and bacteria in the nutrient replete oceans can greatly affect the growth and productivity of the algae (Bell & Mitchell, 1972). In this thesis, I have demonstrated how methods from computational biology can be applied to study and illuminate interactions between algae and bacteria. In each of the three results chapters, a specific set of methods was applied with the aim of uncovering metabolic and genetic aspects central to algae-bacteria mutualism. The studies are based on the model system of the green freshwater alga *L. rostrata* and the soil *alpha-proteobacterium M. loti* (Kazamia et al., 2012b; Grant et al., 2014).

In the Chapter 2, I applied metabolic modelling and flux balance analysis with the aim of studying the metabolism of *M. loti*. The goal of this study was to explore the potential of metabolic modelling as a framework for studying algae-bacterial interactions and the *M. loti* – *L. rostrata* model system. A model for simulating the metabolism of *M. loti* was developed and used to study *M. loti*'s ability to grow on various substrates. I attempted to use the method to identify substrates that could be supplied by *L. rostrata* to *M. loti* when the two organisms are co-cultured.
Ultimately, I was successful in developing a model for simulation of *M. loti* growth and this model could be used to study the effects of genetic insertions and deletions in *M. loti*. While I identified 16 candidate substrates from the flux balance analysis, I was not able to experimentally verify that these were supplied by the algal partner in co-cultures.

As described in the introduction, it has been estimated that approximately half of eukaryotic algae are dependent on exogenous B12 (Croft *et al*, 2005). Bacteria are the only group of organisms that are known to synthesise B12, and the genes related to the synthesis pathway have been isolated and characterised in multiple bacterial species (Warren *et al*, 2002). I therefore set out to identify how common the presence of this pathway is, and which phylogenetic groups of eubacteria are capable of B12 synthesis. The motivation for this study was initially to identify bacteria capable of supplying vitamin B12 to algae, but as the study progressed it was expanded in scope. In addition to the original aim, I studied how common enzymes requiring B12 as a cofactor are in more than 8000 eubacterial species. Among the key findings was that approximately one third of bacteria are capable of B12 synthesis and that this trait is very common in some phylogenetic groups, but very rare in others. I also found that about half of B12-synthesisers are capable of making the lower axial ligand DMB, and therefore cobalamin, and that this is produced by the gene-product of the *bluB* gene in 97% of the cases. By incorporating environmental data about species, I demonstrated that B12 synthesis is found in approximately half of the species in any given environment, except for the rhizoplane, where 80% of the species can produce B12.

In the third research chapter I have described a transcriptomics study of *L. rostrata*. In the study, the gene expression of the alga when grown in axenic conditions and in co-culture with *M. loti* was compared. In this study, the first large scale transcriptome representative for the alga *L. rostrata* was created based on assembly of short reads from an RNA-seq process. The transcriptome for *L. rostrata* will enable rapid hypothesis test related to the model system of *M. loti* and *L. rostrata*. It will be possible to identify the presence of specific genes of interest by performing a simple homology search and primer design for qPCR can take advantage of the available transcriptome (Mořkovský *et al*, 2015). The other main outcome from this study was the identification of 577 contigs.
potentially representing genes that are differentially expressed between the two conditions. From this set of genes, two genes of high interest were identified. The transport enzyme for uptake of the plant hormone auxin was upregulated and the kinase SNRK2.1, which is related to abiotic stress in plants. The remaining contigs still require further analysis and could lead to identification of more mechanisms related to algae-bacterial mutualism.

5.2 Complexity of algal-bacterial interactions

A topic that was encountered multiple times throughout my thesis was the complexity of algal-bacterial interactions. It is possible to identify a metabolite that is being exchanged between a specific alga and a bacterium but based on this and previous studies it appears unlikely that only a single or a few metabolites are involved in the interaction. In the case of the *M. loti* and *L. rostrata* system the observation of increased expression of the auxin uptake transporter (Chapter 4) requires us to consider that more metabolites might be exchanged than the ones initially identified. The model system was developed in a laboratory setting and the two species have not evolved together, since they do not inhabit similar environments: *L. rostrata* is a freshwater alga, and *M. loti* lives in soil (Kazamia et al., 2012b). In our research group we have generated a B12-dependent *C. reinhardtii*, a species that lives in soil and therefore might encounter *M. loti* more frequently (Helliwell et al., 2014a). Nevertheless, the *L. rostrata/M. loti* co-culture is extremely stable (Kazamia et al., 2012b) and exhibits strict regulation (Grant et al., 2014), so they could still have a complex exchange of metabolites. As described in the literature review in Chapter 1 several examples of a simple community of one alga and one bacterium have been shown to exchange multiple metabolites (Bertrand et al., 2015; Amin et al., 2015). The exchanged molecules can also be from within the same class of metabolites such as the two-way exchange of B-vitamins from the study by Cooper et al (manuscript in preparation), which was discussed in Chapter 4. Based on findings from previous studies and from the research presented in this thesis, it appears we are currently only seeing the tip of the iceberg in terms of algae-bacterial metabolite exchanges. Through combinations of genomic, transcriptomic and metabolomics studies more exchanged metabolites are likely to revealed and thus add complexity. So, although modelling efforts, such as the one attempted in this
study, will go some way to capture and simulate the behaviour of these communities, they are unlikely to be accurate reflections until these exchanges are better understood.

5.3 Future studies of algae-bacteria interactions

5.3.1 Using experimental data to validate computational studies
Throughout this thesis, I have aimed to combine experimental and computational data. In experiments, such as qPCR, metabolite measurements and physiological studies it is possible to study a single system or single gene with a high degree of detail. It is also possible to get a sense of reproducibility and variation in the studied system. In computational studies, it is possible to scale in number of studies performed, but it often requires use of a proxy for the hypothesis being tested. An example from this thesis is the presence of genes for B12 synthesis enzymes as a proxy for the ability to carry out B12 synthesis. Using a proxy for the hypothesis testing is sensible when the required data is not available, but it is critical to remind oneself that this is an assumption. It is not a certainty that a bacterium possessing all the required genes for B12 synthesis pathway actually produces B12, even leaving to one side that homologues identified by BLAST are only candidates. It is therefore both sensible and advisable to find a way to confirm this assumption. In the study of B12 synthesis, I did this by searching through the literature for experimental evidence of bacteria producing B12 and then correlating the findings with the computational predictions. By combining experimental and computational data, it is possible to scale up studies from a handful of species to all sequenced species, while maintaining a high degree of trust in the results. While this concept is not new, there is still a lot of computational research being done with limited experimental evidence to validate the method, despite the data being available (Zhang et al, 2009; Degnan et al, 2014a; Peng et al, 2016).

In this thesis, I have demonstrated how a complex pathway and more than 50 genes can be studied simultaneously in more than 8000 bacterial species. By building on the computational framework built here, it will be possible to modify the study to focus on other metabolites and pathways. For a new study, one
would simply have to identify genes related to a pathway or process of interest and collect experimental data that can be used to validate the computationally derived phenotypes. An example of this could be genes related to biotin synthesis and experimental data on biotin synthesis for selected bacteria, or genetic and experimental information about transporters for uptake of an external nutrient.

5.3.2 Auxin and B-vitamins in the context of algae-bacteria interactions

Previous studies have reported that bacteria supply algae with the growth enhancing hormone IAA (an auxin) in ocean communities (Martone et al., 2009; Amin et al., 2015). The increased expression of the auxin uptake transporter in *L. rostrata* suggests that this might also occur in fresh water algae (Chapter 5). IAA synthesis has been studied and described in certain bacteria such as those found in the rhizosphere and the genes encoding the relevant enzymes have been sequenced and characterised. However, it has not been studied how common auxin synthesis is among all bacteria. It would therefore be interesting to undertake a study of IAA synthesis in bacteria using the method and framework developed in chapter 3. This would involve collecting experimental data from published studies on bacteria capable of IAA synthesis as well as gathering representative gene sequences for the relevant enzymes. One aim of such a study would be to test the hypothesis that IAA produced by bacteria is commonly used as a signal molecule in algae-bacterial interactions. This could be done by quantifying the frequency of IAA synthesis among the bacteria commonly associated with algae and comparing it with the frequency among all bacteria. IAA synthesis has been demonstrated among *alpha-, beta- and gamma-proteobacteria* (Duca et al., 2014; De-Bashan et al., 2008). However, it is not known in which phylogenetic groups this trait is common in and which groups it is rare. Using the approach described in chapter 3 it will be possible to test if this trait is limited to proteobacteria or if it is present across all phyla.

In chapter 3 it is described how a bilateral exchange of B-vitamins is potentially occurring in natural environments. In an unpublished study by Cooper et al (in preparation), it was demonstrated that the marine algae *Ostreococcus tauri* and the *alpha-proteobacteria Dinoroseobacter shibae* can form a mutualistic community. In this interaction, *O. tauri* provides niacin, biotin and PABA to the bacteria, while *D. shibae* provides B12. In chapter 3, it was demonstrated how bi-
directional exchange of B-vitamins might be occurring in other pairs of algae and bacteria as well. As the synthesis pathways for niacin, biotin and PABA have been well characterized, it is possible to expand the framework developed to identify bacteria lacking this synthesis pathways. Just as for the proposed auxin study, this expansion should also combine experimental data with the computational data. By finding niacin, biotin and PABA auxotrophs that are able to produce B12, it might possible to find bacteria capable of forming mutualisms with B12-auxotrophic algae.

5.3.3 Using comparative genomics to support synthetic ecology for biotechnology purposes

There has long been an interest in using microalgae for biotechnology purposes such as sustainable biofuel production (Greenwell et al, 2010; Daboussi et al, 2014). Using microalgae for production of low value compounds requires that the process is scaled up from laboratory settings to industrial-sized photobioreactors. One of the challenges in doing this is contamination of the axenic algae cultures from invading bacteria (Grobbelaar, 2012). Creating communities of algae and bacteria has been proposed as way forward to build cultures more resistant to invading species (Kazamia et al, 2012a). Communities of algae and bacteria can be extracted from nature, and attempts have been made to optimize these naturally occurring communities for biotech applications. In a study, aiming to develop an efficient lipid producing culture, a set of closed photobioreactors were inoculated with samples taken from surface waters (Mooij et al, 2013, 2015). To select for high lipid producers, the authors implemented a culturing strategy aimed at selecting for this trait. This involved controlling availability of nitrogen and CO2 and lead to a 5x increase in lipid storage and as high as any reported axenic culture study. Natural communities usually offer advantages such stability, but as they are often complex in the terms of species, it is challenging to apply molecular biology techniques such as gene knockouts or recombinant insertions to these communities. Another approach to building communities is to construct the communities in a laboratory setting from isolated microbial strains and study the productivity of these communities. This last approach has been dubbed synthetic ecology (Dunham, 2007; Kazamia et al, 2012a; Zomorrodi & Segrè, 2016). One of the challenges in synthetic ecology is to select the species for the
community. The selection is likely to be based on a set of desired traits and is thus limited to the species known to possess these traits. In the context of algae-bacterial interactions, it could be desirable to use a B12-producing bacterium in a co-culture with a B12-dependent alga. To expand the list of candidate species, comparative genomics, as demonstrated in chapter 3 could be used to identify bacteria with one or more desired traits. By expanding the group of species with a set of desired traits, a researcher could select species for the community construction based on a wider set of criteria. As an example, it is possible to construct a list of all bacterial species capable of both auxin and B12 synthesis. The species on such a list have the potential to form productive mutualisms with B12-dependent algae.
6 References


Andrews S (2016) FASTQC.


Bauer CB, Fonseca M V, Holden HM, Thoden JB, Thompson TB, Escalante-


Blank LM, Kuepfer L & Sauer U (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* **6**: R49 Available at:


Boursnell C (2016) Transfuse. Available at: https://github.com/cboursnell/transfuse


Brushaber KR, O'Toole GA & Escalante-Semerena JC (1998) CobD novel enzyme with L-threonine-O-3-phosphate decarboxylase activity, is responsible for the synthesis of (R)-1-amino-2-propanol O-2-phosphate, a


De-Bashan LE, Antoun H & Bashan Y (2008) Involvement of indole-3-acetic acid produced by the growth-promoting bacterium Azospirillum spp. in promoting growth of Chlorella vulgaris. *J. Phycol.* **44**: 938–947


Gray MJ & Escalante-Semerena JC (2010) A new pathway for the synthesis of


Jordan A, Gibert I & Barbé J (1994) Cloning and sequencing of the genes from


Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and


Martin JE & Imlay JA (2011) The alternative aerobic ribonucleotide reductase of *Escherichia coli*, NrdEF, is a manganese-dependent enzyme that enables
Available at: http://www.ncbi.nlm.nih.gov/pubmed/21338418


and-salmonella-cellular-and-molecular-biology/oclc/34150023

Nicholls DG aut (2013) Bioenergetics


Pratt R & Johnson E (1964) Production of Pyridoxine and Niacin by Chlorella vulgaris and C. pyrenoidosa. *J. Pharm. Sci.* 53: 151–154 Available at:
http://linkinghub.elsevier.com/retrieve/pii/S0031942207002178
http://www.pnas.org/cgi/doi/10.1073/pnas.1009566107
http://dx.doi.org/10.1016/j.nut.2009.03.010
http://papers2://publication/doi/10.10


J. Bacteriol. 182: 4227–4233 Available at: http://jb.asm.org.proxy-remote.galib.uga.edu/content/182/15/4227.long


van Tol HM, Amin SA & Armbrust EV (2016) Ubiquitous marine bacterium inhibits diatom cell division. ISME J. 1–12 Available at: http://dx.doi.org/10.1038/ismej.2016.112%0Ahttp://www.nature.com/doifinder/10.1038/ismej.2016.112


Vévodová J, Graham RM, Raux E, Schubert HL, Roper DI, Brindley AA, Ian Scott


Available at: http://www.ncbi.nlm.nih.gov/pubmed/20559375


7 Appendix
7.1 Computational framework for discovery of B12 synthesis genes

```python
#!/usr/bin/python -u

def main():
    import sys, os, time
    from Bio import SeqIO
    from Bio.Blast.Applications import NcbiblastpCommandline as blastcmd
    from Bio.Blast import NCBIWWW
    from datetime import datetime
    import threading
    import multiprocessing

    # define blast folder:
    global blast_loc
    blast_loc = '/applications/ncbi-blast+/ncbi-blast-2.2.30+/bin/'

    global sqldb
    sqldb = '/home/ujk20/b12_blaster/sqldb/b12_march16.db'

    global blastdb
    blastdb = '/data/public_data/NCBI/NCBI_nr/nr'

    global profile_db
    profile_db = '/data/public_data/NCBI/NCBI_cdd/cdd_3_13/Cdd.pn'

    global blast_threads
    blast_threads = 1

    max_cpus = 44
    # multiprocessing.cpu_count() - 1

    if len(sys.argv) != 3:
        print 'program exited. Proper usage: taxid gene_file'
        exit()

    print 'running script'

    toplevel = sys.argv[1]
    gene_file = sys.argv[2]

    global folder
    folder = '/home/ujk20/Dropbox/Cambridge/programming/b12_blaster/'
    # folder = os.path.dirname(os.path.realpath(__file__)) + '/temp/
    if not os.path.exists(folder):
        os.makedirs(folder)
```

names_file = '/home/ujk20/taxdump/new_taxid/names.dmp'
tax2name = taxfile2dict(names_file)
toplevel_name = tax2name[toplevel]
print 'Toplevel ' + toplevel_name + '\t' + str(toplevel)

startTime = datetime.now()

tax_file = '/home/ujk20/taxdump/taxdump_220316/nodes.dmp'
(taxids, ranky, parent2children) = chilly(tax_file, toplevel)

gi_tax_file = '/home/ujk20/taxdump/taxdump_220316/gi_taxid_prot.dmp'
#gi_taxid_prot.dmp'
#moso_gi.txt
#p_dentil_gi.dmp
tax2gi_dict = tax2gilist(taxids, gi_tax_file)

print 'Total time: ' + str(datetime.now() - startTime)

import subprocess
reciprocal_blastdb = folder + 'reciprocal_blast_db'
subprocess.call('cp ' + test_gene_file + ' ' + reciprocal_blastdb, shell=True)
subprocess.call('blastloc+makeblastdb -dbtype "prot" -in ' + reciprocal
#/applications/ncbi-blast+/ncbi-blast-2.2.30+/bin/blastp

while not os.path.isfile(blastdb):
    time.sleep(10)
    print 'sleeping'

print 'Working folder: ' + folder

counter = 0
over = 0
threshold = 1000
sql_db = sql_reader(taxids)
#sql_db = {}
(test_gene_list, gene2path) = gene_routine(test_gene_file, folder)

##### To add an extra gene to all the species already in the system
#taxids = set(sql_db.keys())
print 'Tax correcting and file creation started' + 'Total time: ' + str(
(taxids, tax_gene_dict) = tax_corrector(sql_db, taxids, tax2gi_dict, parent;
print 'Tax correcting and file creation done' + 'Total time: ' + str(d:

##### To add an extra gene to all the species already in the system
#taxids = set(sql_db.keys())
print 'Number of species ' + str(len(taxids))
global comb_results
comb_results = {}
from multiprocessing import Process, Queue, Manager
processes = []
indy=0
mgr = Manager()
comb_res2 = mgr.dict()

for taxid in tax_gene_dict.keys():
    #print 'number of active threads ' + str(len(multiprocessing.active_)
    while len(multiprocessing.active_children()) > max_cpus:
        time.sleep(1)
    processes.append((taxid, Process(target=blast_analyser, args=(folder,b1 processes[indy])[1].start())

    indy+=1
    over += 1
    counter +=1
    fraction = max(len(tax_gene_dict)/100,1)
    if counter % fraction == 0:
        fraction = round(float(counter)/ len(tax_gene_dict),2)
        print 'Fraction done ' + str(fraction) + '%
        Total time

for p in processes:
    p[1].join()
    #comb_results = comb_res2
    #print comb_res2
    reskeys = comb_res2.keys()
    for i in reskeys:
        comb_results[i] = comb_res2[i]
        #http://ell.thegreenplace.net/2012/01/16/python-parallelizing-cpu-bound-tasks
        #http://stackoverflow.com/questions/2846653/python-multithreading-for-dummies
        #
        #def f(name):
        #    print 'hello', name
        #
        #if __name__ == '__main__':
        #    p = Process(target=f, args=('bob',))
        #    p.start()
        #    p.join()
        #http://stackoverflow.com/questions/25627313/multiprocessing-dont-use-all
pid = os.getpid()
output = open(folder+'results_'+str(pid)+'.txt','w+')
output2 = open(folder+'sumerized_results'+str(pid)+'.txt','w+')

for tx in comb_results:
    if tx in tax2gi_dict:
        number_of_proteins = len(tax2gi_dict[tx])
    else:
        number_of_proteins = 0
        if tx in tax2name:
            tax2name = tax2name[tx]
        else:
            tax2name = str(tx)

summary = {'aerobic_synthesis_genes':0,'anaerobic_synthesis_genes':0,'anaerobe_not_in':[]}
anaerobe_not_in = []

for item in comb_results[tx].dict:
    for item in comb_results[tx][item]:
        if isinstance(comb_results[tx][item],tuple) and len(comb_results[tx][item]) == 1:
            tub = comb_results[tx][item]

            if tub[0] == 'aerobic_synthesis':
                if tub[3] < 10**-10:
                    summary['aerobic_synthesis_genes'] += 1
                else:
                    aerobe_not_in.append(item)
            elif tub[0] == 'anaerobic_synthesis':
                if tub[3] < 10**-10:
                    summary['anaerobic_synthesis_genes'] += 1
                else:
                    anaerobe_not_in.append(item)
            elif tub[0] == 'B12_related':
                if tub[3] < 10**-10:
                    summary['B12_related'] = 1
                else:
                    summary[item] = 0
            elif tub[0] == 'B12_transport':
                if tub[3] < 10**-10:
                    summary[item] = 1
                    summary['btub_gi'] = tub[4]
                else:
                    summary[item] = 0
            elif tub[0] == 'test':
                summary[item] = tub[3]

        if tx in tax2gi_dict:
if number_of_proteins > threshold:
    # print 'inside here'
    # print nary + 't' + str(summay)
    output2.write(nary + 't' + strl(summay)+'
')
    output.write(nary+'t'+tx+'t'+strl(len(tax2gi_dict[t]))+'
')
else:
    output.write(nary+'t'+tx+'t'+strl(0)+'
')
for item in comb_results[tx]:
    # print comb_results[tx]
    output.write(strl(item) + 't' + strl(comb_results[tx][item])+'
')
output.write('=================================================================='
')
output.close()
output2.close()
# print comb_results
sql_writer(comb_results, tax2name, tax2gi_dict)

print 'Total time: ' + strl(datetime.now() - startTime)
print 'Species/strains analysed ' + strl(over)

def gene_routine(test_gene_file, folder):
    from Bio import SeqIO
    import re, os
    gene_list = []
    gene2path = {}
    for seq_rec in SeqIO.parse(test_gene_file, 'fasta'):
        match = re.search(r'gene:(.+|\|pathway:(.+))', seq_rec.id)
        if match:
            gene = match.group(1)
            pathway = match.group(2)
            gene2path[gene] = pathway
            gene_list.append(gene)
    blastin = folder+str(gene)+'.in.fasta'
    if not os.path.isfile(blastin):
        out = open(blastin, 'w')
        out.write('>'+seq_rec.id+'
'+strl(seq_rec.seq)+'
')
        out.close()
    return (gene_list, gene2path)

def tax_corrector(sql_db, taxids, tax2gi_dict, parent2children, test_gene_list,
    import os
    taxids2 = []
for id in taxids:
    if ranky[id] == 'species' and id in tax2gi_dict:
        if len(tax2gi_dict[id]) > threshold:
            taxids2.append(id)
    else:
        try:
            children = parent2children[id]
        except:
            children = [0]
        temp_dict = {0:0}
        for child in children:
            if child in tax2gi_dict:
                temp_dict[child] = len(tax2gi_dict[child])
            else:
                temp_dict[child] = 0
        key, value = max(temp_dict.items(), key=lambda x:x[1])
        if value > threshold:
            taxids2.append(key)
        else:
            taxids2.append(id)
    elif ranky[id] == 'species':
        taxids2.append(id)

#tax_gene_dict = dict.fromkeys(taxids2,[])
tax_gene_dict = {}
for tt in taxids2:
    if tt in tax2gi_dict:
        if len(tax2gi_dict[tt]) > threshold:
            tax_gene_dict[tt] = []

for taxid in tax_gene_dict:
    if taxid in sql_db:
        genes_in_db = sql_db[taxid]
    else:
        genes_in_db = []
    for test_gene in test_gene_list:
        if test_gene not in genes_in_db:
            tax_gene_dict[taxid].append(test_gene)

in_folder = os.listdir(folder)
files_written = 0
for t in tax_gene_dict:
    if len(tax_gene_dict[t]) == 0:
        del tax_gene_dict[t]
gi_list = t+'gi_list.fasta'
if gi_list not in in_folder:
    files_written +=1
    gi_file = open(folder+gi_list,'w+')
    for gi in tax2gi_dict[t]:
        gi_file.write(str(gi)+'\n')
    gi_file.close()
print 'GI files written '+str(files_written)
return(taxids2,tax_gene_dict)

def sql_reader(taxids):
    import sqlite3 as lite
    import os

    #dir = '/home/ujk20/b12_blaster'
dir = os.path.dirname(os.path.realpath(__file__))
con = lite.connect(sqldb)
prev = {}
with con:
    con.row_factory = lite.Row
    cur = con.cursor()
    cur.execute("SELECT * FROM test")
    rows = cur.fetchall()
    for row in rows:
        tax = str(row['taxid'])
        gene = row['gene']
        if tax in prev:
            prev[tax].append(gene)
        else:
            prev[tax] = [gene]
    return(prev)

def sql_writer(res_dic,taxname,tax2gi_dict):
    import sqlite3 as lite
    import os

dir = os.path.dirname(os.path.realpath(__file__))
#dir = '/home/ujk20/b12_blaster'
print 'farfar dir '+dir
con = lite.connect(sqldb)
headers = ('id int,Type Text, Sci_name TEXT, num_prot int, pathway te')
with con:
    cur = con.cursor()
    for tx in res_dic:
if tx in tax2name:
    sciname = tax2name[tx]
else:
    sciname = tx
if tx in tax2gi_dict:
    num_prot = len(tax2gi_dict[tx])
else:
    num_prot = 0
tubby = (tx,sciname,num_prot,'NA',0,0,float(100))
# print tubby
continue
for gene in res_dic[tx]:
    # print gene + '\t' + str(res_dic[tx][gene])
    pathway = res_dic[tx][gene][0]
    coverage = round(float(res_dic[tx][gene][1]),5)
    identities = round(float(res_dic[tx][gene][2]),5)
    eval = float(res_dic[tx][gene][3])
    gi = res_dic[tx][gene][4]
    reciprocal_gene = str(res_dic[tx][gene][5])
    domains_list = str(res_dic[tx][gene][6])
    # print 'reci far fafr' + reciprocal_gene + 'farfar'
tubby = (tx,sciname,num_prot,gene,pathway,coverage,identities,
    # print tubby
    #(taxid int, Sci_name TEXT, num_prot int, gene text ,pathway 1
    cur.execute("INSERT INTO test VALUES "+str(tubby))

def blast_analyser(folder, blastdb, taxid, test_gene_list, gene2path, comb_res):
    # gene_list = tax_gene_dict[taxid]
    import Bio.Blast.Applications
    import Bio.Blast.NCBIXML
    import cStringIO
    import re
    import subprocess

    results = {}  
    for gene in test_gene_list:
        resline = ()
        reciprocal_gene = "NA"
        pathway = gene2path[gene]
        results[gene] = (pathway,0,0,100,00000000)
        blastin = folder+str(gene)+".in.fasta"
        blastout = folder+str(gene)+"_blast_out.fasta"
        gi_list = folder+taxid+".gi_list.fasta"
stdout, stderr = blastp_cline()
blast_xml = cStringIO.StringIO(stdout)

for l in blast_xml:
    print l

blast_record = Bio.Blast.NCBIXML.read(blast_xml)
query_length = blast_record.query_length
if len(blast_record.alignments) > 0:
    for align in blast_record.alignments:
        hsp = align.hsps[0]
        coverage = 0
        coverage = float(hsp.align_length) / max(float(query_length),

if match:
    gi = str(match.group(1))
else:
    gi = 'NA'
break

target_file = folder + "target_seq" + taxid + "".fasta
get_entry_cmd = blast_loc + "blastdbcmd -db " + blastdb + " -dbtype output = subprocess.call(get_entry_cmd, shell=True)
for reciprocal_blast_gene in reciprocal_blast(target_file, reciprocal_list):
    domains_list = domain_finder(target_file, gene)
    resline = (pathway, coverage, float(hsp.identities))/float(hsp.alnlen)
def domain_finder(target_file, gene):
stdout, stderr = blastp_cline()
blast_xml = cStringIO.StringIO(stdout)
blast_record = Bio.Blast.NCBIXML.read(blast_xml)
prev = 0
domains_list = []
domain = ""

eval = 10000

if len(blast_record.alignments) > 0:
    for align in blast_record.alignments:
        for hsp in align.hsps:
            if hsp.score <= prev:
                continue
            match = re.search('\s(.+?),',align.title)
            if match:
                domain = match.group(1)
                eval = hsp.expect
                domains_list.append((domain,eval))

return (domains_list)

def reciprocal_blast(target_file,reciprocal_blastdb):
    #Reciprocal Matches serach

    #Finding the reciprocal result
    blastp_cline()

    blast_xml = cStringIO.StringIO(blastp_output)
    blast_record = Bio.Blast.NCBIWWW.read(blast_xml)
    prev = 0

    if len(blast_record.alignments) > 0:
        for align in blast_record.alignments:
            for hsp in align.hsps:
                if hsp.score <= prev:
                    continue
                match = re.search('gene:(.+)?/pathway',align.title)
                if match:
                    gene_reciprocal = match.group(1)
                    return(gene_reciprocal)
                else:
                    print 'no match - Something is wrong in reciprocal m:

return('NA')

def tax2prot(txid,out_file):
    import os.path
    from Bio import SeqIO
    #if os.path.isfile(out_file):
        #print 'Already downloaded genome'
        #return(0)

    from Bio import Entrez
import os
import sys
Entrez.email = 'ujk20@cam.ac.uk'
searchhandle = Entrez.esearch(db='protein',term=str('txid' + taxid),retmc
searchresults = Entrez.read(searchhandle)
protids = searchresults['IdList']
len_prot = len(protids)

out = open(out_file, 'w+')
for prot in protids:
   out.write(prot + '
')
out.close()
return(len_prot)

def chilly(dump_file, id):
dump = open(dump_file, 'rU')
parent2children = {}
ranky = {}
for line in dump:
   line_list = line.split('|')
   child = str(line_list[0]).strip()
   parent = str(line_list[1]).strip()
   rank = str(line_list[2]).strip()
   ranky[child] = rank
   if parent in parent2children:
      parent2children[parent].append(child)
   else:
      parent2children[parent] = [child]

if id in parent2children:
   children = parent2children[id]
else:
   return([id], ranky, parent2children)
for chi in children:
   if chi in parent2children:
      children.extend(parent2children[chi])
children.append(id)
end_points = []
for chi in children:
   if ranky[chi] == 'species':
      end_points.append(chi)
   if chi in parent2children:
      end_points.extend(parent2children[chi])
if ranks[id] == 'species':
    end_points.append(id)
return(set(end_points), ranks, parent2children)
def tax2gi_list(taxids, gi_tax_file):
    tax2gi_dict = {}
    file = open(gi_tax_file, 'rU')
    for line in file:
        line_list = line.split()
        tax = line_list[1]
        gi = line_list[0]
        if tax not in taxids:
            continue
        if tax in tax2gi_dict:
            tax2gi_dict[tax].append(gi)
        else:
            tax2gi_dict[tax] = [gi]
    for tax in taxids:
        if tax not in tax2gi_dict:
            tax2gi_dict[tax] = []
    return(tax2gi_dict)
def taxfile2dict(filename):
    import sys, re
    file = open(filename, 'rU')
    size_dict = {}
    dict = {}
    for line in file:
        match = re.search('scientific name', line)
        if match:
            ary = line.split(' |')
            dict[ary[0].strip()] = ary[1].strip()
    return(dict)
if __name__ == '__main__':
    main()
7.2 Cyanobacteria and Eukaryotic Algae Use Different Chemical Variants of Vitamin B12
Cyanobacteria and Eukaryotic Algae Use Different Chemical Variants of Vitamin B₁₂

Graphical Abstract

Complex B₁₂ cycling in the photic zone

Pseudocobalamin producers

Cyanobacteria

Remodelling algae

PsCbl

Non-modelling algae

Cbl

DMB

Cobalamin producers

Highlights

- Dominant marine cyanobacteria synthesize only pseudocobalamin
- Pseudocobalamin is orders of magnitude less bioavailable to eukaryotic algae
- Certain algae can remodel pseudocobalamin to a bioavailable form
- This implies a complex B₁₂ cycle between microbes in the photic zone

Authors

Katherine Emma Helliwell, Andrew David Lawrence, Andre Holzer, ..., David John Scanlan, Martin James Warren, Alison Gail Smith

Correspondence

katherine.helliwell@mba.ac.uk (K.E.H.), as25@cam.ac.uk (A.G.S.)

In Brief

Helliwell et al. demonstrate that the two major groups of photosynthetic microbes in the photic zone, cyanobacteria and microalgae, use different forms of vitamin B₁₂. These findings challenge the oversimplified assumption of a linear flux of B₁₂ from producers to consumers, in favor of a more complex network of B₁₂ production, uptake, and cycling.

Helliwell et al., 2016, Current Biology 26, 1–10
April 25, 2016 ©2016 The Authors
http://dx.doi.org/10.1016/j.cub.2016.02.041
Cyanobacteria and Eukaryotic Algae Use Different Chemical Variants of Vitamin B₁₂

Katherine Emma Helliwell,²,³,⁴,⁵ Andrew David Lawrence,³ Andrea Holzer,³,⁴ Ulrich Johan Kurahl,¹ Severin Sassa,²,⁶ Bernhard Kräutler,⁷ David John Scanlan,⁵ Martin James Warren,² and Alison Gail Smith¹

¹Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK
²School of Biosciences, University of Kent, Kent CT2 7NJ, UK
³Institute of Pharmacy and Molecular Biotechnology, Ruprecht-Karls University Heidelberg, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany
⁴Institute of Organic Chemistry and Centre of Molecular Biosciences, University of Innsbruck, Innrain 58/1b, 6020 Innsbruck, Austria
⁵School of Life Sciences, University of Warwick, Coventry CV4 7AL, UK
⁶Present address: Marine Biological Association of the UK, Citadel Hill, Plymouth PL1 2PB, UK
⁷Correspondence: katherine.helliwell@mba.ac.uk (K.E.H.), azs25@cam.ac.uk (A.G.S.)

SUMMARY

Eukaryotic microalgal and prokaryotic cyanobacteria are the major components of the phytoplankton. Determining factors that govern growth of these primary producers, and how they interact, is therefore essential to understanding aquatic ecosystem productivity. Over half of microbial species representing marine and freshwater habitats require for growth the corrinoid cofactor B₁₂, which is synthesized de novo only by certain prokaryotes, including the majority of cyanobacteria. There are several chemical variants of B₁₂, which are not necessarily functionally interchangeable. Cobalamin, the form bioavailable to humans, has as its lower axial ligand 5,6-dimethylbenzimidazole (DMB). Here, we show that the abundant marine cyanobacterium Synechococcus synthesizes only pseudocobalamin, in which the lower axial ligand is adenine. Moreover, bioinformatic searches of over 100 sequenced cyanobacterial genomes for B₁₂ biosynthesis genes, including those involved in nucleotide loop assembly, suggest this is the form synthesized by cyanobacteria more broadly. We further demonstrate that pseudocobalamin is several orders of magnitude less bioavailable than cobalamin to several B₁₂-dependent microalgae representing diverse lineages. This indicates that the two major phytoplankton groups use a different B₁₂ currency. However, in an intriguing twist, some microalgal species can use pseudocobalamin if DMB is provided, suggesting that they are able to remodel the cofactor, whereas Synechococcus cannot. This species-specific attribute implicates algal remodelers as novel and keystone players of the B₁₂ cycle, transforming our perception of the dynamics and complexity of the flux of this nutrient in aquatic ecosystems.

INTRODUCTION

Eukaryotic microalgae are photosynthetic microbes estimated to be responsible for up to 50% of global carbon fixation [1]. Elucidation of factors that control algal community structure and dynamics is thus fundamental to understanding the global cycling of carbon. Nutrients such as iron, nitrogen, and phosphorus clearly play an important role [2], but many microalgae also require the vitamins B₁ (thiamine), B₆ (biotin), or B₁₂ for growth [3]. B₁₂ is required as a cofactor for methionine synthase (METH: EC 2.1.1.13) activity, a key enzyme of cellular one-carbon (C1) metabolism important for production of the universal methyl donor S-adenosylmethionine (SAM), and for folate cycling necessary for DNA synthesis [4]. Those algae that do not need a supply of B₁₂ cannot synthesize the vitamin; rather, they possess an alternative form of methionine synthase (METE: EC 2.1.1.14) that can catalyze the same reaction in a B₁₂-independent fashion [5–7]. Measurement of B₁₂ levels in the water column have indicated concentrations of ~10 pmol/L for freshwater ecosystems [7] and are often below the threshold of detection in certain marine habitats, including large areas of the northeast Pacific margin [8]. The scarcity of this micronutrient is therefore thought to limit phytoplankton abundance [8], so competition for B₁₂ among those organisms that require it is likely. Indeed, field-enrichment experiments found that, whereas N addition stimulated all microbial growth, there was a specific growth enhancement of phytoplankton >5 μm (i.e., the larger eukaryote fraction) with B₁₂ supplementation [9]. However, several recent studies have demonstrated that heterotrophic bacteria can satisfy microbial requirements for B₁₂ via mutualistic interactions (e.g., [10]).

Provision from prokaryotes is particularly pertinent because the biosynthetic pathway of this cofactor is confined to certain archaea and bacteria. B₁₂ is an umbrella term that refers to cobalt-containing corrinoids (ring-contracted tetrapyrroles), which have upper and lower axial ligands to the cobalt ion (Figure 1A). The nature of these ligands varies, leading to diversity in the structural forms of B₁₂. Methylcobalamin, where the upper axial ligand is a methyl group, is involved in methyl-transfer
Figure 1. Synechococcus Strains Synthesize Pseudocobalamin Rather Than Cobalamin

(A) Structural variants of B12: with AdoB12 (coenzyme B12); with DMB; with cyanocobalamin (CN-Cbl); with methylcobalamin (Me-Cbl); with 5,6-dimethylbenzimidazole (DMB); with adenosylcobalamin (AdoCbl); with methylcobalamin (Me-Cbl); and with pseudocobalamin (Pse-Cbl). Cobalamin is therefore considered not “bioavailable” to humans, and the efficacy of vitamin supplements produced from cyanobacteria such as Spirulina and Aphanothece, which also contain pseudocobalamin, has been questioned [15, 16].

Cyanobacteria are the numerically dominant photosynthetic microbes in marine environment [17]. Two cyanobacterial strains, Crocosphaera watsonii WH8501 and Synechococcus sp. WH7803, were reported to release B12 into the media at rates exceeding those estimated for the heterotrophic bacterium Halomonas, suggesting that cyanobacteria might be the major source of B12 for marine algae [18]. However, indications from the early literature suggest consideration of algal specificity toward different B12-like factors may be pertinent [18]. Here, we investigate corrinoids in several strains of Synechococcus, an abundant and ubiquitous marine cyanobacterium [17, 20], the nature of their axial ligands, and their ability to support growth of eukaryotic primary producers.

RESULTS

Synechococcus Species Make Only Pseudocobalamin

The biosynthesis of the corrinoid ring of B12 from the common tetrapyrrole progenitor uroporphyrinogen III requires at least 20 enzymatic steps, and several routes are known [21]. In a preliminary investigation, Saffudo-Wilhelmy et al. [22] searched for the presence of B12-biosynthesis genes in ~40 marine cyanobacteria with sequenced genomes. They found that all but one species had putative homologs for at least 11 of these genes and so concluded that they were capable of making B12. However, they did not investigate the genes involved in synthesis of the lower axial ligand and so could not conclude whether the cyanobacteria studied produced cobalamin or pseudocobalamin. To refine the analysis of B12-biosynthesis genes, we searched 123 sequenced cyanobacterial genomes for all 20 genes of the corrinoid pathway. All but six species contained at least 10/20 of these genes and were predicted to be B12 producers [Data S1; Supplemental Experimental Procedures]. Additionally, we searched for genes involved in DMB biosynthesis, for which two routes are currently known [23, 24]. The Blus enzyme, first characterized in Sinorhizobium meliloti (Rhizobia) makes DMB from riboflavin under aerobic
conditions [23]. Using this sequence as a query, no hits were found in 118 cyanobacterial genomes including C. watsonii WH8101 and Synechococcus sp. WH7803. For five species (including three from the *Flosculentus* genus), hits for Blub were obtained: two were annotated as cob(yl)ytic acid a-c-diamide reductase (CoDp), which is an enzyme of an earlier stage of B**6** biosynthesis, whereas the others were unknown. In contrast, Blub homologs were found in 80% of sequenced rhizobia (227/284 genomes; Data S1C) and 65% of Rhodococcus species (77/128; Data S1D) including *Mesorhizobium loti*, *Sino-

Brucellum meliloti*, *Rhizobium leguminosarum*, and the marine bacterium *Dromonebacter shibahae*, all of which can support algal *B**6**-*autotrophic growth [20, 21]. More recently, a second route for DMB biosynthesis was identified in the obligate anaerobic bacterium *Eubacterium limosum* [24], and enzymes encoded by the *baxABGCE* operon were shown to direct DMB production via an oxygen-sensitive reaction from the purine precursor 5-aminooimidazole ribotide (AIR) [24]. We found that none of the cyanobacterial genomes encoded the full *baxABGCE* operon (Data S1B). Moreover, CoDp, which is required for DMB activation, is absent from all but two cyanobacterial genomes. Taken together, these searches suggest that the vast majority of cyanobacteria cannot make DMB.

To validate the observations from the bioinformatics analysis, we wanted to assess directly what carinoids are synthesized by cyanobacteria so we investigated the **B**2 content of strains of marine *Synechococcus*, because this is an ancient and ecologically abundant lineage [25] with a mean global abundance of 7.0 ± 0.3 x 10^6 cells L^-1, high-biomass-specific CO2 fixation rates [17, 26] and axenic strains are available. Carinoids can be extracted from cells as their carino-derivatives and then analyzed by high-performance liquid chromatography (HPLC) mass spectrometry (LC-MS). First, using purified cyanobacterial carinoid (obtained commercially) and cyanobacteriocobalamin, prepared from Propionibacterium acidipropionic (12), we were able to distinguish the two varieties of **B**2 by their different retention times on the LC (Figure 1B) and different mass (Figures S1A and S1B). Derivatized cell lysates obtained from axenic cultures of the heterotrophic marine bacterium *D. shibahae* DFL127 contained only cyanocobalamin. We next tested five members of the *Synechococcus* lineage representing different clades and habitats (highlighted in red in Figure S2): coastal strain CCS311 (sub-cluster [SQ] 5.1A; clade I); oligotrophic WH8102 (SQ5.1A; clade II); WH7803 and WH7805 (SQ5.1B; clades V and VI, respectively) which are widely distributed in various oceanic waters; and the eutrophic strain WH7621 (SQ5.2). A single peak was observable in these samples at a retention time consistent with the pseudocobalamin standard (Figure 1B), and its identity was confirmed by MS (Figures S1D-S1H). To facilitate subsequent physiological work, we also tested two model freshwater cyanobacterial species *Synechocystis* sp. PCC6803 and *Synechococcus elongatus* PCC7942, because these species grow quickly and easily in the laboratory. Again, cell lysates from these strains contained only pseudocobalamin (Figures 1B, S1L, and S1J). Together, these data demonstrate that the *Synechococcus* species sampled here make only *pseudocobalamin* in axenic laboratory culture conditions.

However, some **B**2-synthesizing bacteria can modify endogenous **B**2 forms with an alternative base [27]. For instance, although Salmonella enterica cannot make DMB, it can import it and then make cobalamin instead of pseudocobalamin [27]. To investigate whether *Synechococcus* can perform this so-called "guided biosynthesis," we grew strains WH8102 and WH7803 (representing SQ5.1A and 5.1B) in the presence of pseudocobalamin and 1 μM DMB, but only pseudocobalamin was detected (Figures 1B, S1K, and S1L). We conclude therefore that *Synechococcus* cannot replace the adenine base with DMB to make cobalamin.

**Pseudocobalamin Is Orders of Magnitude Less Bioavailable to Eukaryotic Algae** We next tested whether cyanobacterially derived **B**2 could be utilized by eukaryotic algae. Cell-free extracts of *S. elongatus* PCC7942 (pseudocobalamin producer) were unable to rescue growth of the *B**2**-dependent freshwater alga *Lobomonas rostrata*, whereas there is clear growth with the addition of extracts of three rhizobial bacteria (Figure S3), which all encode Blub [28]. It is conceivable that the growth is due to other compounds in the crude lysate, so this initial experiment was extended using the purified compounds, cyanocobalamin and cyanobacteriocobalamin. Equivalent concentrations of each **B**2 variant were added to axenic cultures of *B**2**-dependent microalgae from different algal lineages: marine species *Ostreococcus tauri* (*Chlorophyta, Mamiillitophyceae*); *Amphidinium carterae* (Alveolata, Dinophagidaceae); *Pavlova lutheri* (*Haptophyta, Pyrrhophyceae*); *Thalas-
sisoria pseudorosana* (*Heterokontophyta, Coscinodiscophyceae*); *Aureococcus anophagefferens* (*Heterokontophyta, Pelagophyceae*); and the freshwater species *Euglena gracilis* (*Euglenata, Euglenozoa*) and *L. rostrata* (*Chlorophyta, Volvocales*). We also tested a *B**2**-dependent merl* mutant of *Chlamydomonas reinhardtii* (*Chlorophyta*, Volvocales) [29]. When pseudocobalamin was supplied at a concentration of 0.07 or 0.7 μM, we observed little or no growth in any of the marine species, nor with the *C. reinhardtii* merl* mutant or *L. rostrata*. This is in contrast to cobalamin, which supported growth of all algal cultures at equivalent concentrations (Figure 2; Student’s t test; p < 0.05, n = 3). For *O. tauri*, *A. carterae*, and *T. pseudorosana*, and to a lesser extent the *C. reinhardtii* merl* mutant, provision of pseudocobalamin at 7 μM (~10 μg/L) supported growth to a similar extent as cobalamin (Figures 2A, 2B, 2D, and 2E), although this amount is significantly higher than found in natural ecosystems (with reported concentrations ranging from below the detection threshold to 0.03 μM across large areas of the northeast Pacific margin, for instance [7, 8]). One way to compare the efficacy of the different **B**2 variants is to carry out dose-response experiments, which enable determination of an EC50 (that is, the effective concentration required to support half-maximal biomass accumulation) and also provide an indication of the minimum cobalamin concentration required to support growth, and so we carried these out with the *C. reinhardtii* merl* mutant. Figure 2G shows that the EC50 was ~0.07 μM for cobalamin, compared to ~7 μM (~100-fold higher) for pseudocobalamin. In addition, it is clear that even the highest concentration of pseudocobalamin used (40 μM) is not saturating, whereas 0.2 μM cobalamin supports maximum growth. For *E. gracilis*, some growth was observed even at the lowest pseudocobalamin concentration, but it was still significantly lower than with cobalamin (Figure 2G). Thus, pseudocobalamin is orders of magnitude less...
bioavailable to eukaryotic algae. It is notable to mention that *E. gracilis* has also been demonstrated to encode a B12-dependent (type I) ribonucleotide reductase [35], which could account for the growth response to pseudocobalamin observed in this alga.

We reasoned that the reduced ability of pseudocobalamin to support growth of algal B12-autotrophs may be either because the molecule cannot be used as a cofactor or because it does not get transported into algal cells. To investigate the latter possibility, we took advantage of the presence of B12-responsive genes previously identified in the marine diatom *Phaeodactylum tricornutum* and *C. reinhardtii* [31, 33]. These algae do not need B12 for growth but will uptake and use it if it is available [1-3]. Several genes in these algae are responsive to B12: METE (in *P. tricornutum* and *C. reinhardtii*) [6, 31, 33]; CBA1, encoding a novel cobalamin acquisition protein (in *P. tricornutum* only) [17]; and S-adenosylhomocysteine hydrolase, SAHY (in *C. reinhardtii* only) [33]. Using qRT-PCR, we analyzed their expression in cells grown in the presence of 0.7 nM cobalamin or pseudocobalamin. For *P. tricornutum*, both forms of B12 resulted in downregulation of METE, but the effect was less pronounced with pseudocobalamin compared to that with cobalamin (Student’s *t* test; *p* < 0.001; *n* = 3; Figure 3A). As previously demonstrated, cobalamin suppressed CBA1 [31], but this gene was significantly upregulated by pseudocobalamin (Student’s *t* test; *p* < 0.0001; *n* = 3; Figure 3A). In *C. reinhardtii*, both METE (Student’s *t* test; *p* < 0.05; *n* = 3) and SAHY (Student’s *t* test; *p* < 0.01; *n* = 3) were downregulated relative to the no supplementation control (Figure 3B) with both forms of B12. Subsequent western blot analysis using polyclonal antibodies against *C. reinhardtii* METE protein [33] demonstrated a modest reduction of METE abundance when cells were grown with pseudocobalamin, although not to the same extent as with cobalamin. Nonetheless, the effect of pseudocobalamin on the expression of these four B12-responsive genes indicates that the molecule can enter both *C. reinhardtii* and *P. tricornutum* cells.

**Certain Algae Are Capable of Remodeling Pseudocobalamin**

By analogy with the guided biosynthesis described earlier, some bacteria that cannot synthesize B12 de novo can modify imported forms via “remodeling” [33, 34]. To investigate this possibility in algae, we grew B12-requiring species in the presence of pseudocobalamin and a range of DMS concentrations. For most growth, no DMS was added to DMB supplement (Figures 4A, 4B, and 4D-4G). However, for *P. lutheri* and the *C. reinhardtii* metaII mutant, addition of DMS alongside pseudocobalamin rescued growth to the same extent as cobalamin (Figures 4C and 4H). A dose-response experiment with *P. lutheri* established an *EC*50 value of ~18 pM for cobalamin (Figure 5A). A similar experiment with
a fixed concentration (0.7 mM) of pseudocobalamin but varying the amount of DMB revealed a similar EC_{50} (~23 mM; Figure 1B). Interestingly, an equivalent dose-response curve (and EC_{50} value: ~26 mM) was observed when cells were grown in medium made using natural filtered seawater rather than artificial sea salts. Thus, the level of DMB in the natural filtered seawater is not sufficient to allow remodeling; otherwise, the dose-response curve would be shifted to the left. That comparable levels of B_{12} and DMB (at a fixed level of pseudocobalamin) are able to rescue B_{12}-dependent growth implies that P. lutheri is remodeling pseudocobalamin with DMB to generate cobalamin. Dose experiments with the C. reinhardtii B_{12}-dependent metE mutant also identified similar EC_{50} values of ~28 mM and ~70 mM for cobalamin and DMB, respectively (Figures 1C and 5B). In this case, the EC_{50} value for DMB was slightly higher than cobalamin. We also tested whether C. reinhardtii is capable of de novo lower-loop synthesis and grew the C. reinhardtii B_{12}-dependent metE mutant with DMB alongside di(3cyano)cobalamin, a B_{12} precursor that lacks the DMB ribonucleotide tail, but no restoration of growth was observed (Figure 1A).

Our data indicate that, of eight diverse algal species studied, six do not appear to be able to use exogenous DMB. Nevertheless, the observation that growth of the C. reinhardtii B_{12}- dependent metE mutant (alongside that of P. lutheri) with pseudocobalamin is restored by DMB provision suggests that these algae are able to chemically modify pseudocobalamin to a form that can support B_{12}-auxotrophic growth. To test more directly whether pseudocobalamin is being remodeled, we grew samples of C. reinhardtii in the presence of (1) cobalamin, (2) pseudocobalamin (1 mM), and (3) pseudocobalamin (1 mM) + DMB (1 mM) and prepared cell lysate for LC-MS analysis. However, we could not detect any form of B_{12} from lysed cells. We infer from this that intracellular B_{12} levels are extremely low; i.e., the quantity from ~1 x 10^{12} cells is below the threshold detection of the LC-MS (which in our system is ~1 x 10^{-7} molecule). Without a clear idea of what order of magnitude more biomass would be required and constrained by the limitations of scale, we turned to alternative means of characterizing remodelling activity and investigated the effect of DMB + pseudocobalamin on gene expression in C. reinhardtii. Previously, we had generated several transgenic lines of C. reinhardtii expressing the B_{12}-responsive element of the METE gene fused to the BLE gene, which confers resistance to the antibiotic Zeocin [32]. This reporter gene construct enables rapid and easy measurement of B_{12}-responsive gene expression, whereby growth with cobalamin represses expression of BLE so that cells die in the presence of Zeocin (Figure 1B). In contrast, pseudocobalamin alone had little effect, but the inclusion of DMB impaired growth to the same extent as cobalamin, demonstrating that C. reinhardtii converts DMB and pseudocobalamin into a form that is able to repress the METE promoter.

The pathway for pseudocobalamin remodeling has been investigated previously in the purple bacterium Rhodobacter sphaeroides, and cobalamin amidohydrolase (CobZ) and cobalamin-phosphate synthase (CobE) have been implicated in this process [14]. We could not identify CsbZ or CsbB in any of the algal genomes we analyzed. Therefore, we searched for proteins shown to be involved in lower-loop assembly and activation [15] in E. coli (CobT, CobI, and CobG) [32], where mutants of CobT are unable to incorporate exogenous DMB (Figure 6D). We identified genes encoding all three of these proteins in C. reinhardtii, which exhibits the remodeling phenotype (Table S1; Figure 6G). In contrast, BLASTP searches of the genomes of O. tauri and T. pseudonana, species that do not appear to remodel, were negative for CobT and CobS. Although a hit for CobT in C. reinhardtii is reported to be present in T. pseudonana, it should be noted that CobC catalyzes a dephosphorylation step [21, 27], and therefore BLAST searches may retrieve genes encoding unrelated phosphatases. Interestingly, we identified hits for CobT and CobS, but not CobA, in A. anophagefferens, which can use pseudocobalamin with DMB, but only with very high levels of the latter (10 mM; Figure 6B). Transcript sequences for P. lutheri are available via the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP), a database of 398 unique strains representing ecologically significant and taxonomically diverse marine microbial eukaryotes [35]. This alga, which can use pseudocobalamin alongside DMB, also expresses CObET, CObIS, and CObGC (Data S1A). Thus, the presence of these novel proteins correlates with the ability to remodel pseudocobalamin, implicating them in B_{12} metabolism. We also identified another 46 candidate remodellers (Data S1A) including several that encoded METE, and so we believe C. reinhardtii is being independent of a source of B_{12} for growth. In total, the potential remodellers include representatives of the higher class levels Alkedastra, Stramenopila, Hacrobia, and Virdiplantae (Data S1B). Incidentally, none of the sequenced Synechococcus genomes encode CobT (Data S1), which might explain why Synechococcus strains cannot remodel pseudocobalamin to cobalamin in the presence of DMB (Figure 1B).
Figure 4. Provision of Lower Ligand Substrate DMB Together with Pseudocobalamin Can Support Growth of Certain B12-Dependent Algae (A–H) Species were grown in liquid medium (Table S2) without or with 0.7 mM cobalamin (open bars) or 0.7 mM pseudocobalamin (shaded bars) in the presence of different DMB concentrations in batch culture over several transfers or until the cells died in the B12 treatment. (A) O. tauri (OT985), (B) A. carteri, (C) P. lutheri, (D) T. pseudonana, (E) A. anophagefferens, (F) L. rostrata, (G) E. gracilis, and (H) C. reinhardtii. B12-dependent evolved (red) mutant lines are shown [9]. Optical density (OD675) was used to quantify growth (mean ± SEM; n = 3).

that pseudocobalamin is the major form of B12 synthesized by most if not all cyanobacteria.

We found that pseudocobalamin is considerably less bioavailable than cobalamin to several B12-dependent algae (Figure 2). This reduced bioavailability suggests these organisms are compromised in their ability to acquire or use pseudocobalamin as a cofactor. Human intrinsic factor, part of the B12 uptake system in the gut, exhibits a 500-fold-lower binding affinity for pseudocobalamin [14], thus reducing the bioavailability of the compound to humans. In algae, currently only one protein has been implicated in B12 uptake [27] (CBA1), although the precise molecular mechanism and role of CBA1 in B12 binding are not fully understood. Nevertheless, the ability of pseudocobalamin to affect the expression of algal B12-responsive genes (Figure 3) and protein levels (Figure S5) suggests this compound can enter algal cells, albeit that it has the opposite effect on CBA1 to cobalamin, suggesting that the cells are experiencing cobalamin deficiency. Transport of pseudocobalamin into the cell is also indicated by our observed remodeling of pseudocobalamin in C. reinhardtii and P. lutheri following DMB addition (Figures 4 and 5). The identification of genes encoding enzymes of lower ligand activation (COBIT) and nucleotide-loop assembly (COBS) [17] in these algae (Table S1) provides a likely mechanism for cornoid remodeling. We found no evidence of secretory peptide signals in C. reinhardtii COBIT or COBS using the green algal subcellular localization tool “PredAlgo” [29], implying that remodeling takes place within the cell and providing further support for the ability of pseudocobalamin to be taken up. Whether these genes have been acquired through lateral gene transfer from a bacterial source, which is thought to be the case for E. coli [37], remains unknown. However, of the non-algal sequences retrieved via a BLAST search of the

DISCUSSION

Eukaryotic microalgae and cyanobacteria are the major components of the phytoplankton in marine and freshwater systems. Because they both inhabit the photic zone, they will compete for resources including light and limiting nutrients such as nitrogen and Fe. We have demonstrated that, in contrast to heterotrophic bacteria such as D. shibae (and certain rhizobial bacteria) [13] that make cobalamin, members of the ubiquitous marine Synechococcus genus synthesize only pseudocobalamin, in which the lower base is adenine instead of DMB (Figure 1). Moreover, a survey of diverse cyanobacterial genomes, encompassing marine and freshwater species, showed the vast majority do not encode bluB or the bluABCDE operon (Table S1) [23, 24]. This strongly suggests
NCBI non-redundant database with the C. reinhardtii COBS/CORB top hits were derived from the amphipod-associated prokaryotic species Aphanoxyces ataci (46–57) and Sphaeroforma arctica (76–77), respectively. A broader phylogenetic analysis of COB genes will be integral to further understanding of what, at a first glance, appears to be an intriguing evolutionary history. Because algae rely on B12 for METH [1], the function of pseudocobalamin as a cofactor for this enzyme is also an important question. Structural data available for the B12-binding pocket and the active site of METH [39, 40] implicate several amino acids in B12 binding, with the DMB “tail” buried within a cleft of the active site [40]. Because pseudocobalamin contains an alternative lower base to B12, it seems plausible that algal METH proteins may have reduced binding affinity for pseudocobalamin. The combination of DMB with pseudocobalamin improves the biocatalytic to certain algae. We infer from this that these remodeling algae are able to generate cobalamin from pseudocobalamin + DMB, although we were unable to measure detectable levels of any form of B12 in C. reinhardtii cells grown under these conditions. It is possible therefore that another form of the vitamin is being generated, though we deem this unlikely. In any case, our results highlight the importance of considering environmental concentrations of DMB. A bioassay to measure free DMB concentrations was recently reported [41]. Analysis of samples derived from host-associated (bacteria/rumen) and

Figure 5. Certain Algae Can Remodel the Lower Axial Ligand of Pseudocobalamin with Exogenously Supplied DMB. (A and B) Relative growth yield of (A) P. lutheri cells supplemented with different concentrations of cobalamin after 19 days (values of OD650 were normalized as a proportion of growth at 1000 μM B12 or 85 P. lutheri cells supplemented with different concentrations of DMB in the presence of 0.7 μM pseudocobalamin after 19 days in artificial seawater or natural filtered seawater (values of OD650 normalized to growth at 1000 μM DMB). (C) and (D) Equivalent experiments with C. reinhardtii, evolved B12-dependent methyl mutant [46] are displayed in (C) and (D) after 96 h growth (mean ± SEM; n = 3).

Figure 6. Further Characterization of the Remodeling Phenotype. (A) Relative growth yield (OD650) of C. reinhardtii B12-dependent mutant grown with DMB alongside (dicyanocobamide, a B12 precursor that lacks the DMB nucleotide tail). Cells were grown in liquid medium (Table S1) with 0.7 mM cobalamin (open bars), 0.7 μM pseudocobalamin (gray bars), or 0.7 μM dicyanocobamide (black bars) in the presence of different DMB concentrations. (B) Relative growth yield (OD650) of C. reinhardtii reporter line containing a Zoc1 resistance gene controlled by the MET1 promoter [22] after 13 days in the presence (white bar) or absence (black bar) of DMB (1 μM) and 20 μg/ml Zeocin without or with 0.7 μM cobalamin or pseudocobalamin. Values of OD650 were normalized as a proportion of growth with no B12 (mean ± SEM; n = 3). (C) The pathway for the activation of DMB and nucleotide loop assembly in S. enterica (adapted from [73]). CoB-T catalyzes the attachment of a phosphoribosyl moiety derived from nicotinate mononucleotide to form e-nicotinamide phosphate. CobG and CobC catalyze the attachment of the activated base to the cobamide precursor (GDP-cobamide).
and refine strategies of B12 acquisition/utilization in order to enhance accessibility to this limiting micronutrient is strong. In any case, the importance of B12 and its derivatives in structuring microbial communities in aquatic ecosystems may have been previously underestimated.

**EXPERIMENTAL PROCEDURES**

**Bioinformatics Approaches**
A full description of sequence similarity search parameters is provided in the Supplemental Experimental Procedures.

**Chemicals**
Upper axial cyan forms of cobalamin/pseudocobalamin were used for all B12-amendment experiments. Cyanocobalamin was purchased from Sigma-Aldrich, UK. Cyanopseudocobalamin was prepared by guided biosynthesis from a culture of Pseudomonas aeruginosa DSM 20273 as described previously and confirmed by UV-Vis, circular dichroism (CD), mass, and nuclear magnetic resonance (NMR) spectroscopic analysis [15].

**Strains and Growth Conditions**
Details of microbial strains and culture conditions are provided in the Supplemental Experimental Procedures and Table S8.

**Molecular Methods**
RNA Extraction and qRT-PCR
Total RNA was extracted [16] and treated with the Ambion Turbo DNase-Free Kit to remove genomic DNA. RNA was reverse transcribed into cDNA with SuperScript II (Invitrogen). Details of qRT-PCR are given in the Supplemental Experimental Procedures and Table S3.

**Western Blotting**
Total protein was extracted and western blot experiments performed as described in [16].

**ACCESSION NUMBERS**
The data described in this manuscript are available at https://www.repository.cam.ac.uk/handle/1810/65413.

**SUPPLEMENTAL INFORMATION**
Supplemental information includes Supplemental Experimental Procedures, five figures, three tables, and two datasets and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.02.041.

**AUTHOR CONTRIBUTIONS**

**ACKNOWLEDGMENTS**
We thank G. Wheeler (MBA, UK) for providing L4 seawater. We also thank M. Croft, S. Hosen, M. Scully, L. Norman, and C. Kaczmara for technical support and helpful insight. Funding was from the BBSRC BB/I033164/1 to K.E.N. and A.G.S. and BMBF0086941/1 to A.D.L. and M.J.W. EU FP7 Marie Curie ITN PhotoComm, nos. 317184 to A.G.S. and U.J.K., The Swiss National Science Foundation (grant nos. PBEZA-115703 and P400PS-124169 to S.S.), and the EU project NaGUMBA (no. 319795 to D.J.S.).

Received: January 4, 2016
Revised: February 15, 2016
Accepted: February 17, 2016
Published: March 31, 2016

8 Current Biology 26, 1–10, April 25, 2016
REFERENCES


new subcellular localization prediction tool dedicated to green algae. Mol.
38. Drennan, C.L., Huang, S., Drummond, J.T., Matthews, R.G., and
40. Tolstoguzov, M., Künzel, R., Hilden, B.H., Marsh, E.N.G., and
dynamics of the B12-binding subunit of glutamate mutase from Clostridium tetra-
morphum. Structure 6, 1021-1033.
for the detection of biotinylated proteins reveals the presence of a range of
environmental samples. Front Microbio 5, 992.
42. Haal, K.R., Carlson, L.T., Devol, A.H., Ambroset, E.V., Moffett, J.W.,
and other B vitamins in seawater by liquid chromatography/tandem mass
43. Men, Y., Seth, E.C., Yi, S., Crofts, T.S., Allen, R.H., Tapa, M.E., and
modification in dechlorinating microbial communities. Environ.
Microbio 17, 4872-4884.
structure depends on both lower ligand availability and CoBT substrate
7.3 Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis
Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta)

Susan H. Brawley1,4, Nicolas A. Blouin2,4, Elizabeth Ficko-Blean1, Glen L. Wheeler4, Martin Lohn4, Holly V. Goodson2, Jerry W. Jenkins2,4, Cristen L. Blaby-Haas1, Katherine E. Helliwell1, Cheong Xin Chan1, Tara N. Marriage1, Debashish Bhattacharya1, Anita S. Klein1, Yacine Badia1, Juliet Brodie2, Yuanyu Cao2, Jonas Collé1, Simon M. Dittami1, Claire M. M. Gachon6, Beverley R. Green1, Steven J. Karpowicz4, Jay W. Kim1, Ulrich Johann Kudahl7, Senjie Lin1, Gurvan Michel, Maria Mittag1, Bradley J. S. C. Olson4, Jasmin L. Pangilinan1, Yi Peng1, Huan Qiu1, Shengqiang Shu1, John T. Singer1, Alison G. Smith1, Brittanly N. Sprecher1, Volker Wagner1, Wenfeng Wang1, Zhi-Yong Wang1, Juaying Yan1, Charles Ziesemer1, Simonne Zúñiga-Altukhov1,3, Yufan Zhuang3,1,2, Kerri W. Barry1, Daniel S. Roksar3, Jeremy Schmutz2,5, John W. Stiller2, Arthur R. Grossman1, and Simon E. Prochnik1

1School of Marine Sciences, University of Maine, Orono, ME 04469. 2Department of Molecular Biology, University of Wyoming, Laramie, WY 82071; 3Sophonics University, University Pierre and Marie Curie Paris 06, CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 9074, 29688 Roscoff, France. 4Marine Biological Association of the United Kingdom, Plymouth, PL1 2BQ, United Kingdom. 5Institute for Molecular Physiology, Pflanzenbiochemie, Johannes Gutenberg-Universität Mainz, 55128 Mainz, Germany. 6Department of Chemistry & Biochemistry, University of Notre Dame, South Bend, IN 46556. 7HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806. 8Department of Energy, Joint Genome Institute Walnut Creek, CA 94596. 9Biology Department, Brookhaven National Laboratory, Upton, NY 11973. 10Department of Plant Sciences, University of Cambridge, Cambridge CB2 1FA, United Kingdom. 11Institute for Molecular Biology, University of Queensland, Brisbane, QLD 4072, Australia. 12School of Chemistry and Molecular Biology, University of Queensland, Brisbane, QLD 4072, Australia. 13Division of Biology, Kansas State University, Manhattan, KS 66506. 14Department of Ecology, Evolution & Natural Resources, Rutgers University, New Brunswick, NJ 08901. 15Department of Biological Sciences, University of New Hampshire, Durham, NH 03824. 16Scottish Association for Marine Sciences, Scottish Marine Institute, Oban PA37 1QA, United Kingdom. 17Natural History Museum, Department of Life Sciences, London SW7 5BD, United Kingdom. 18Department of Botany, University of British Columbia, Vancouver BC, Canada V6T 1Z4. 19Tau Biosciences LLC, Edison, NJ 08803. 20Department of Biomedical Engineering, University of California, Santa Cruz, CA 95064. 21Department of Marine Sciences, University of Connecticut, Groton, CT 06340. 22Institut für Allgemeine Botanik und Pflanzenphysiologie, Friedrich-Schiller-Universität Jena, 07743 Jena, Germany. 23Department of Molecular & Biomedical Sciences, University of Maine, Orono, ME 04469; 24Bush Forestry and Proteomics Research Center, HSTI, Fujian Agriculture and Forestry University, Fuzhou 350002, China. 25Department of Plant Science, Carnegie Institution for Science, Stanford, CA 94305. 26Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT 06269. 27Institute of Molecular Physiology and Biotechnology, University of Bern, 3011 Bern, Switzerland, and 28Department of Biology, East Carolina University, Greenville, NC 27834

Edited by Stephen R. Palumbi, Stanford University, Pacific Grove, CA, and approved June 6, 2017 (received for review February 22, 2017)

Porphyra umbilicalis (laver) belongs to an ancient group of red algae (Bangiophyceae), is harvested for human food, and thrives in the harsh conditions of the upper intertidal zone. Here we present the 87.7-Mbp haploid Porphyra genome (95.8% G + C content, 13.12% gene loci) and elucidate traits that inform our understanding of the biology of red algae as one of the few multicellular eukaryotic lineages. Novel features of the Porphyra genome shared by other red algae relate to the cytoskeleton, calcium signaling, the cell cycle, and stress-tolerance mechanisms including photosynthesis. Cytoskeletal motor proteins in Porphyra are restricted to a small set of kinases that appear to be the only universal cytoskeletal structures within the red algae. Dynamin-type GTPases are absent, and most red algae, including Porphyra, lack myosin. This surprisingly minimal cytoskeleton offers a potential explanation for why red algal cells and multicellular structures are more limited in size than in most multicellular lineages. Additional discoveries further relate to the novel stress-tolerance of bangiophytes including ancestral enzymes for sulfation of the hydrophilic galactan-rich cell wall, evidence for mannan synthesis that originated before the divergence of green and red algae, and a high capacity for nutrient uptake. Our analyses provide a comprehensive understanding of red algae, which are both commercially important and played a major role in the evolution of other algal groups through secondary endosymbioses.

cytoskeleton | calcium-signaling | carbohydrate-active enzymes | stress tolerance | vitamin B12

T he red algae are one of the founding groups of photosynthetic eukaryotes (Archeplastida) and among the few multicellular lineages within Eukarya. A red algal plastid, acquired through secondary endosymbiosis, supports carbon fixation, fatty acid synthesis, and other metabolic needs in many other algal groups in ways that are consequential. For example, diatoms and lepadophytes have strong biogeochemical effects; apicomplexans cause human disease (e.g., malaria); and dinoflagellates include both coral symbionts and toxigenic producing “red tides” (1). The evolutionary processes that produced the Archaeplastida and secondary algal lineages remain under investigation (2–5), but it is clear that both nuclear and plastid genes from the ancestral red algal have contributed dramatically to broader eukaryotic evolution and diversity. Consequently, the imprint of red algal metabolism on the Earth’s climate system, aquatic foodwebs, and
Significance

Fossil evidence shows that red algae (Rhodophyta) are one of the most ancient multicellular lineages. Their ecological, evolutionary, and commercial importance notwithstanding, few red algal nuclear genomes have been sequenced. Our analysis of the Porphyra umbilicalis genome provides insights into how this macrophyte thrives in the stressful intertidal zone and into the basis for its nutritional value as human food. Many of the novel traits (e.g., cytoktotic organization, calcium signaling pathways) we find encoded in the Porphyra genome are extended to other red algal genomes, and our unexpected findings offer a potential explanation for why the red alga are constrained to small stature relative to other multicellular lineages.

Human health is immense. Moreover, the oldest taxonomically resolved multicellular eukaryote in the fossil record (1.2 Ga) is the benthic red alga Bangiomorpha, which closely resembles the extant marine alga Bangia (6). As typical of many bioniques, Porphyra grows in one of Earth’s most physically stressful habitats, the intertidal zone, where organisms are exposed to daily and seasonally fluctuating temperatures, high levels of irradiance (including UV), and severe osmotic stress and desiccation. Porphyra and its ancestors have competed successfully in this dynamic and severe environment for over a billion years, through numerous changes in climate and mass extinctions. Here we describe the genome of Porphyra umbilicalis. Examination of the Porphyra genome and complete genomes of other red algae (Chondrus crispus (7), Cyanidioschyzon merolae (8), Galderia sulphuraria (9), Porphyridium purpureum (10), Pyropia yezoensis (11)) revealed numerous additional differences between the red algae and other eukaryotic lineages, including a reduced complement of motor proteins, unique signaling molecules, and augmented stress tolerance mechanisms, especially in Porphyra.

Results and Discussion

Genomic Analysis. An 87.7-Mbp assembly of the P. umbilicalis (hereafter, Porphyra) nuclear genome was generated from RAD (restriction-site associated DNA) sequencing, with insertions and deletions corrected using Illumina whole-genome shotgun reads (SI Appendix, Methods). The Porphyra genome has a substantial red content (90%) for a compact genome, with the most common repeat classes being DNA (15.5 Mbp) and LTR (14.9 Mbp) elements (SI Appendix, Table S5). Gene models were predicted at 13,125 loci using de novo gene prediction algorithms supported by evidence from protein homology and expression data (SI Appendix, Table S6). A typical gene has ~2 exons, implying abundant splicing for a red alga; however, only ~225 alternative splice-forms were identified from expressed sequence tag coverage of genes (SI Appendix, Table S6). Overall, the genome is 65.8% G+C, but protein-coding regions average 72.9% and reach up to 94% G+C (SI Appendix, Fig. S7). Nearly 98% of the sequenced transcripts (expressed sequence tags) can be mapped to the genome assembly, and we identified complete complements of genes encoding RNA polymerase subunits and all other conserved proteins involved in transcription, translation, and DNA synthesis (SI Appendix, Table S10), suggesting that the genome is nearly complete.

Phylogenomic analysis (12) of the red algae (Rhodophyta) distinguishes a class (Cyanidiophyceae) of extremophilic unicellular species and two sister classes of mesophilic species, which we refer to here as the SCRP (Stylonematothecaceae, Compsogonophyceae, Rhodophyceae, Porphyridiophyceae) and the BF (Bangiophyceae, Florideophycaceae) (SI Appendix, Fig. S11). The SCRP clade contains unicells, microscopic filaments, and microscopic blades, whereas the BF clade holds macrophytes (“seaweeds”) that comprise the majority of described species (13). Phylogenomic comparisons of Bangiophyceae (e.g., *P. umbilicalis*, *P. yezoensis*) and Florideophyceae (e.g., *C. crispus*, *Callithamnion tuberculatum*) suggest that these two red algal classes are highly diverged (14). The absence of some pathways and genes from red algae is likely because of genomic reduction in the red algal ancestor (4), and we confirmed that Porphyra lacks genes described previously as lost in other red algae, including those encoding enzymes of the glycosyl-phosphatidylinositol (GPI) anchor biosynthesis pathway (Kyoto Encyclopedia of Genes and Genomes (KEGG), 22 genes), autophagy proteins (KO pathway ko04140, 17 genes), and most flagellar proteins (4).

Cytoskeleton. The cytoskeleton of red algae is poorly characterized, despite the long-recognized absence of flagella from the red algae (14, 15). Nuclear-associated organelles that lack counterparts appear to organize the mitotic spindle (1), and freeze-substitution reveals cytosplastic microtubules and bundles of actin microfilaments (16). Cytoskeletal inhibitors and fluorescent probes (e.g., FITC-phalloidin for microfilaments) demonstrate that actin microfilaments form cortical rings during cytokinesis and during sperm/egg fusion, ensheath migrating secretory vesicles and organelles, and are prominently labeled in amoeboid red algal spores (17-21). Certainly the composition of the red algal cytoskeleton must determine many of the capabilities and limitations of red algae because of the fundamental roles the cytoskeleton plays in intracellular transport, secretion of cell wall materials, regulation of cell size and shape, and responses to developmental and environmental signals that influence cell polarity and complex tissue development in many eukaryotes (22, 23). Here we report that Porphyra and other red algae have significantly reduced cytoskeletons and consider the consequences for size and complexity.

We identified four closely related actin genes in Porphyra, as well as the chromatin remodeling, actin-related protein (ARP), and myosin (4) nuclear-associated organelles that lack counterparts appear to organize the mitotic spindle (1), and freeze-substitution reveals cytosplastic microtubules and bundles of actin microfilaments (16). Cytoskeletal inhibitors and fluorescent probes (e.g., FITC-phalloidin for microfilaments) demonstrate (21). How these spores move with their ARPs is an intriguing question; perhaps they rapidly polymerize microfilaments with the aid of other nucleating machinery, such as formins (22), which are present. However, although many animal cells and plants are members of expanded family genes that have been differentiated to support processes required to build complex morphologies (e.g., polarized tip growth and cell plate orientation) (25, 26), Porphyra has only two formins (M Appendix, Table S15). In addition to formins, Porphyra and other red algae (M Appendix, Table S16) contain profilin, which interacts with formin-cofilin, a key depolymerizing factor; and severin, which cuts microfilaments to promote remodeling. However, we did not find other well-conserved, widely distributed actin-modifying proteins (e.g., WASP/WAVE, CapZ, fimbrin) in Porphyra, and few if any convincing homologs in other red algae (M Appendix, Table S16). The most striking limitation to microfilament-mediated phenomena in Porphyra and most other red algae is the absence of myosin.

Myosin genes were not detected in any of the available genomes of the BF clade (Fig. 1), and all nonspecific Myo inhibitors (23) were used previously (e.g., ref. 21) to infer myosin activity. We do find that the single myosin annotated previously (9) in the extremophilic Galderia, but absent from *Cyanidioschyzon* (8), is also found in three classes of the SCRP clade, but not in Porphyridiophyceae (Fig. 1 and M Appendix, Fig. S17).
We found that the Porphyrin genome encodes the expected α- and β-tubulin proteins, as well as some proteins related to tubulin folding (e.g., Porphyrin contains cofactors B and D, but not A or E) and microtubule nucleation (e.g., γ-tubulin, γ-complex proteins) (SI Appendix, Tables S15 and S16). However, many of the expected tubulin regulatory proteins that are widely distributed across eukaryotes and even found in other red algae are missing. For example, Porphyrin contains EB1 and Mort1/XMAP215, two highly conserved proteins of the MT plus-end tracking (+TIP) complex, but the +TIP CLASP and the cross-linker MAP506 appear absent, even though these highly conserved genes are present in other red algae (SI Appendix, Table S16). Unknown as yet is whether Porphyrin simply lacks these activities, or has recruited other proteins to fill their roles. Porphyrin and other red algae lost flagellar and cytoplasmic dynein motors, intermediate chains, and most light chains. A dynein heavy chain reported from C. curvispina might be from a contaminant or horizontal gene transfer (SI Appendix, Table S16). Porphyrin and other red algae retain a particular light chain that is also conserved in flowering plants, which independently lost flagellar motility (SI Appendix, Table S16). This dynein light chain is expressed under abiotic stress and phytohormone treatments in plants (28); thus, it may have a role in the stress tolerance of Porphyrin. In contrast to the loss of the dynein motor, Porphyrin does have representatives of several kinesin motor subfamilies, specifically kinesins 5, 7, and 14, which are expected to be involved in spindle assembly, kinetochore function, and regulation of microtubule dynamics, respectively (Fig. 1 and SI Appendix, Fig. S18 and Tables S15 and S16). Porphyrin also contains three divergent kinesins that usually group with the mitotic motor kinesin 13. However, classic vesicle transport motors (kinesin subfamilies 1, 2, and 3) appear to be absent in Porphyrin, which is particularly surprising considering the apparent loss of myosins and dynesins.

Given the surprising paucity of motors, how do Porphyrin cells accomplish intracellular transport of membranes or other cargo? One answer may be provided by the observation that the "mitotic" motor kinesin 14 (present in all sequenced red algae, see Fig. 1) can act as a minus-end directed transporter in land plants (reviewed by refs. 29 and 30). In addition, a few red algae do contain members of the kinesin 4 subfamily (Fig. 1), which are reported to be plus-end directed microtubule vesicle motors in plants (31). However, Porphyrin and most other sequenced red algae lack this protein. Thus, unless kinesin 5 or kinesin 17 has unexpected functionality, Porphyrin and the majority of red algae (BF clade) would also apparently lack a plus-end directed kinesin-microtubule motor system. Taken together, these data show that the paucity of motors explains the absence of cell streaming from red algae including Porphyrin.

One counterpoint to the frequent absence of near-ubiquitous cytoskeletal proteins is that Porphyrin does have two septins (SI Appendix, Tables S15 and S16). Filament-forming proteins that are involved in cytokinesis, cell polarity, and membrane remodeling (22, 23). Regardless, our overall analysis of Porphyrin and other red algae with sequenced nuclear genomes [Chondrus (7), Cystodeschyzon (8), Galliera (9), Porphyra (10), Pyropia (11)] indicates that the red algal cytoskeleton lacks the complexity and diversity of cytoskeletal elements present in other multicellular lineages (Fig. 1). Although it is possible that regulatory proteins are simply too divergent to recognize, the paucity of motors is especially apparent. We suggest that this observation could help to explain long-standing questions about morphological evolution in the red algae, including the lack of parenchyma in these organisms. Compared with other multicellular lineages (green algae/plants, brown algae, fungi, animals), the abilities to form large cells and large multicellular structures appear to be limited in the red algae. For example, the largest cells in Porphyra are found in its holdfast, which is composed of thousands of thread-like, slow-growing rhizoid cells that are millimeters long, and some species of Grifithia (32), which is a subtidal florideophyte, have cells ~2-mm long. In contrast, large cells filling special niches or functions evolved in multiple freshwater and marine green algae, including eooeetids (1), germinating pollen tubes in land plants (33), as sporangiospores in fungi (34), as nerve cells in animals (35), and as sieve elements in brown algae (36). Maintenance of large cells would be expected...
to require vigorous multidirectional intracellular transport, which seems unlikely with a motor repertoire as limited as that seen in Porphyra. A 12-kb 20-gene cluster of red algae with sequenced genomes (Fig. 1).

Similarly, brown algae (46-kb kelps) (36), animals, and plants assemble large, complex 3D body plans with true parenchyma, but multicellular forms of red algae mostly consist of simple filaments or filaments interwoven and tacked together by secondary pit plugs (1) (i.e., pseudoparenchyma). Red algae are usually ≤50 mm long and only a few species reach 2 m in length (36, 37). Fungi cannot make parenchyma, but saprophytic mycelia can form large multicellular structures that is not straightforwardly attributed to the complex extracellular matrix; however, plants, which retain two mitotic gene sets (myosins, kinesins) despite their independent loss of dynein, suffer serious stunting and other developmental abnormalities following gene knockouts of myosins (25, 39). Analysis of the mitochondrial transcripts and developmental regulators does not offer a compelling explanation for why red algae have failed to evolve tissues comparable to those in brown algae, plants, and animals (40). Taken together, our comparative analysis of the genomes of Porphyra and other red algae leads us to speculate that the small number of cytokinetic elements in red algae compared with those in other multicellular lineages (Fig. 1) has constrained the ability of red algae to develop larger, more complex cells and multicellular structures.

**Stress.** The ecological success of Porphyra and many of the closely related benthic macroalgae (37, 41) in the intertidal zone suggests that these species developed cellular mechanisms to cope with this harsh environment. In particular, Porphyra grows from the mid-to-high intertidal zone, where it is routinely exposed during daily low tides to light, desiccation, and extreme fluctuations in temperature and acidity. Bladders can lose up to 95% of their water on some days, but are metabolically active as soon as they are rehydrated by the rising tide (24). Here we infer novel adaptations to cope with these stresses.

**Phytophthora.** Light is required for photosynthesis, but severe cellular damage can be caused by exposure to phototoxic organics to the high levels of light (visible and UV) that are present in mid-to-high intertidal zone where most benthic phycian algae, including Porphyra, grow. Porphyra has the same complement of genes (SI Appendix, Table S19) to carry out photophosphorylation and electron transport (ETC) 1 and 2, which are typically involved in photic regulation of the photosynthetic antenna during the shift from darkness to light (45). Porphyra also has 11 genes encoding "high-light-in-duced" or "one-hit" proteins (here ORPs), which are involved in photoinhibition of the photosynthetic electron transport chain, and cell viability under stressful environmental conditions (43, 44). Mechanistically, ORPs may regulate chlorophyll and tetrapyrrole biosynthesis, stabilize photosystem 1 (PSI), bind free chlorophyll or chlorophyll breakdown products from damaged PSI complexes during the damage/repair cycle, or bind carotenoids that dissipate excess absorbed light energy. In contrast to the 11 Porphyra ORPs (SI Appendix, Table S19), we found 4 ORPs in Chondrus, which experiences less drying and light stress because of its low intertidal/sublittoral habitat, 6 in P. yezoensis, 7 in Porphyridium, and only 1 ORP in Cyanidioschyzon, which inhabits a stable hot spring environment (SI Appendix, Photosynthesis, Phytophthora, Stress Genes). More analysis is needed, but the putative gene family expansion in Porphyra suggests positive selection for increased gene dosage. Porphyra was one of the first organisms where quenching of excess excitation energy in response to desiccation stress was observed (44), but the molecular mechanisms responsible for this quenching in red algae remain unclear.

Porphyra encodes genes for catalases and peroxidases, as well as the biosynthesis of numerous antioxidants, such as ascorbic acid (vitamin C) (8) and lycopene (Fig. S20). When overexpression of photosynthetic electron transport occurs and reactive oxygen is generated, catalase detoxifies hydrogen peroxide in red algae, and the expansion of a cata
calase gene family in Porphyra and *Pyropia* (five genes) (SI Appendix, Table S21) compared with other red algae (one to two genes) could reflect the demand for detoxification of reactive oxygen species that cannot diffuse away from blades exposed by the falling tide to high light and air while they are still hydrated and photosynthetically active. Toopherox prevent photooxidative damage of polyunsaturated fatty acids (45), and the *Porphyra* tetraphenolic metabolite antenna (46, 47). The 32 heat shock proteins (Hsp) in *Porphyra* indicate a possible expansion of the Hsp10 family (SI Appendix, Table S20), which are chaperones of Hsp70 and play an important role in protein maturation and repair under normal and stressed conditions (48, 49).

Porphyra is frequently exposed to elevated intensities of UV radiation in the intertidal zone and shows remarkable tolerance to both UV-A and UV-B (50, 51). Porphyra has at least two strategies to protect photosynthesis and other key cellular processes from UV damage: mycosporine-like amino acids (MAAs) and circadian control over the timing of UV-sensitive processes.

MAAs act as "sunscreens" and comprise up to 1% of the dry weight of Porphyra, with the compound porphyra-334 being the major MAA (51). Four proteins—MysA, MysB, MysC, and MysD—are required to synthesize these MAAs in anoxygenic phototrophic cyanobacteria, such as *Nostoc* (52, 53), whereas MysD is replaced by a non-photosynthetic peptide synthase in *Anabaena* (Fig. 2). Cyanobacterial MysD shows a relaxed substrate specificity, with condensation of threonine instead of serine onto mycosporine-glycine to yield porphyra-334 (53). The Porphyra genome contains a gene cluster that codes for the fusion protein FpsD (54) (Fig. 2 and SI Appendix, Figs. S23 and S24). The presence of the MysD gene in *Porphyra* and the occurrence of UV-B events suggest that this arrangement provides a selective advantage and efficient MAA biosynthesis for red algae that experience high UV irradiances.

Developmental and abiotic stress responses are often associated with photoceptors in eukaryotes. The plant circadian clock contains blue- and red-light photoceptors, including cryptochromes (CRY) and phytochromes (PHY), to entrain the circadian clock (55, 36); these photoceptors are also involved in other fundamental processes in plants, including growth and development. *Porphyra* does not appear to encode a PHY photoreceptor or a typical plant CRY photoreceptor, although it has maintained four genes of the CRY/photolyase family. The *Porphyra* CRY that most like plant CRYs is similar to a DNA
photolyase (PHR2) (SI Appendix, Fig. S27). In contrast, other red algae seem to encode plant CRYs; however, these group closely to the cyclobutane pyrimidine dimer class III photolyases (58) in some cases (SI Appendix, Fig. S27). Porphyrin also encodes a class II cyclobutane pyrimidine dimer photolyase (Pum00220035.1) as well as a CRY-DASH protein (59) (Pum02440001.1) in a separate subfamily of CRYs that displays DNA repair activity; these proteins are present from bacteria to vertebrates (59). In addition, there is an animal-like CRY (Pum04010002.1) encoded on the Porphyrin genome (SI Appendix, Fig. S27). Animal-like CRYs could function as blue-light photoreceptors associated with the entrainment of the circadian clock, as they do in Drosophila and some other insects, or could even add an oscillatory component to the clock, as in mouse or humans (59). The animal-like CRY in Porphyrin is closely affiliated with the cryptochrome photolyase family (e.g., Ostracococcus tauroi, Phaeocystis tricornis in SI Appendix, Fig. S27), which has maintained photolyase activity (60, 61), in contrast to other plant and animal-like CRYs. Moreover, cryptochrome photolyase family CRY can affect transcriptional activity in a heterologous clock system (e.g., mammalian CRY) and control blue-light-dependent cellular processes, as found with insect or plant CRYs (60, 61). Similar to the Olaminobacter reinhardtii animal-like CRY (aCRY), the Pum04010002.1 protein could sense blue light and possibly other light qualities, including red light (62, 63); this red-light sensing ability stems from the formation of the neutral radical form of the photoreceptor’s flavin chromophore.

**Signaling and Homeostasis.** Porphyrin must cope with significant osmotic and ionic stress in the intertidal zone. Several low molecular weight carbohydrates act as compatible solutes in red algae, including Porphyrin (64–66). We found enzymes encoded on the Porphyrin genome that support floridoside and iso-floridoside synthesis, but no evidence for digenase synthesis (SI Appendix, Table S31). Turain is a likely osmolyte in Porphyrin (67), and we found homology of enzymes, such as cysteine dioxygenase, implicated in metazoan turain biosynthesis, which suggests an ancient origin for eukaryotic biosynthesis of this amino acid derivative (SI Appendix, Table S31). Porphyrin also has a wide range of ion transporters, including an unusual class of Na+/H+ exchangers that are most similar to the NhaA class of transporters from α-Proteobacteria and are distinct from the Na+/H+ exchangers previously identified in eukaryotes. Porphyrin has two P2C-like Na+/K+ ATPases and a P3A H+/ATPase similar to those in land plants, suggesting Porphyrin is able to energize its plasma membrane with either Na+ or H+ for secondary active transport, which could aid its survival in the intertidal zone (SI Appendix, Table S31). The Porphyrin genome also contains a range of Ca2+-permeable membrane channels that could play a role in osmotic stress signaling, including two homologs of OSCA (SI Appendix, Table S31), a recently identified Ca2+ channel in land plants (68); however, the Ca2+ sensor kinases through which Porphyrin senses and responds to cytosolic Ca2+ elevations appear to be distinct from land plants and, indeed, from those of all other eukaryotes.

Land plants possess two expanded families of Ca2+ sensor kinases, the Ca2+-dependent protein kinases (CDPKs) and the calcium B-like protein (CBL)-interacting protein kinases (CIPKs), which are activated through the binding of CBL (69, 70). The CIPKs have been characterized extensively in the green lineage (Viridiplantae), and they are present in other eukaryotes, including stramenopiles, haptophytes, and excavates, suggesting a likely origin early in eukaryotic evolution (71). Surprisingly, we found that genes encoding CIPKs and CBLs are absent from the Porphyrin genome and from all other available red algal genomes and transcriptomes. The CDPKs are also absent from Porphyrin and notably from other haptophytes and floridosiphones, but they are present in red algae in the SCR5 clade (Fig. 3; compare SI Appendix, Fig. S11 and Table S32). Another class of Ca2+ sensor kinases, the Ca2+/calmodulin-dependent protein kinases (CAMKs), which are important in both plants and animals, are also missing in Porphyrin. This finding suggests that much of the extensive network of Ca2+ sensor kinases found in other eukaryotes is absent in the red algae (Fig. 3). Whereas the Porphyrin genome encodes several proteins with domains similar
Fig. 3. Module structure and presence or absence of different Ca²⁺-sensor kinase in Porphyra and other eukaryotic CAMK, CPIK, CDPK, CDTKL. The CDTKLs from Porphyra and other red algae represent a newly recognized family of protein kinases belonging to the TKL family that contains multiple Ca²⁺-binding EF hands.

The CDTKLs from Porphyra and other red algae represent a newly recognized family of protein kinases belonging to the TKL family that contains multiple Ca²⁺-binding EF hands. Instead, we found that Porphyra possesses a class of Ca²⁺-sensor kinases with two to three Ca²⁺-binding EF-hand domains at the N terminus and a kinase domain at the C terminus (Fig. 3 and SI Appendix, Tables S32 and S33). The kinase domain from all known Ca²⁺ sensor kinases belongs to the CAMK group of kinases, whereas this uncharacterized protein belongs to the tyrosine kinase-like (TKL) class of kinases (72, 73). The genes encoding homologous proteins are present on many other red algal genomes, including Chondrus, Cyanidioschyzon, and Galdieria, but appear to be absent from all other characterized eukaryotic genomes. Hence, the Ca²⁺-dependent TKLs (CDTKLs) represent a newly recognized class of Ca²⁺-regulated kinases that appear to be unique to red algae (Fig. 3).

The sucrose nonfermenting-1-related kinase (SnRK) family of serine/threonine kinases plays important roles at the interface between metabolic and stress signaling from fungi (yeast) to land plants (74). There are 38 SnRKKs divided into three subclasses (SnRK1, -2, -3) in Arabidopsis, and the Porphyra genome encodes 31 proteins with homology to the kinase domains of Arabidopsis SnRKS proteins (SI Appendix, Fig. S34). Phylogenetic analysis places the Porphyra SnRK group in two clades that include Arabidopsis SnRK1 and SnRK3/CPK, and a third expanded clade containing 19 unique members (SI Appendix, Fig. S34). Although no Porphyra SnRKS are most similar to the SnRK2 subclass of Arabidopsis, several members of the SnRK3/CPK clade contain motifs similar to the C-terminal domain E3 of SnRK2 (SI Appendix, Fig. S34), which mediates interaction with PP2C in the abscisic acid (ABA) signaling pathway (75, 76). Furthermore, there are homologs of several ABA biosynthetic genes in Porphyra (SI Appendix, Fig. S34), and ABA synthesis and responses to exogenous ABA are reported for Porphyra umbilicalis (77). Although highly speculative, the evidence raises the possibility that the SnRK family plays important roles in stress responses including ABA-mediated responses in Porphyra. Genomic data also support the presence of ethylene-mediated regulation in Porphyra, because the Porphyra genome contains genes encoding several proteins involved in ethylene biosynthesis (SI Appendix, Fig. S35), and ethylene was detected and shown to induce stress responses in red algae (78).

The cell cycle is regulated by dimers of cyclin and cyclin-dependent kinases (CDKs). Except in C. merolae, where circular rhythms and stress responses regulate the G1/S transition (79), the red algal cell cycle is not well characterized. Results here, coupled with the earlier studies of C. merolae, show that regulation of the red algal cell cycle is similar to that of metazoans and plants (80, 81), except that we did not find cyclin D encoded in any red alga, including Porphyra (SI Appendix, Table S37). Cyclin Ds are well conserved in all eukaryotes where they regulate the G1/S transition. The histone-specific cyclin A (CYCA), a known cell-cycle progression regulator, might instead function in place of D-type cyclins in red algae. We found that the glucophytes (Archaeplastida) Cyanophora parasita (2) also lacks a cyclin D homolog.

Plant Defense Genes. Bangiophytes pathogen include some oomycetes, viruses, and bacteria, and these organisms sometimes cause serious economic losses to nori aquaculture. Land plants detect pathogens via an array of cell surface pattern-recognition receptors (e.g., receptor-like kinases) and intracellular receptors (e.g., nucleotide-binding domain and leucine-rich repeat peptide)-mediated responses (82). We found no evidence for this higher plant type of pathogen detection, but two families of intracellular ligand-binding protein containing NB-ARC domains, and a family of potential intracellular receptors containing malectin and Ig-like fold-domains, were found in the multicellular red algal genomes (SI Appendix, Fig. S38). We also found a bangiophyta-specific family of at least three proteins harboring WFA and C-type lectin domains (SI Appendix, Table S39). Ctype lectin domains are involved in pathogen detection in some animals (83, 84).

Cell Walls. Red algae synthesize many unique polysaccharides, such as again that contribute to their ecological success and are of significant interest to biotechnological and industrial applications. The cell wall of blade cells is not cellulose but partially crystalline β-1,4-linked mannin in the outer cell wall, with a highly crystalline inner layer of β-1,3-linked xylan (85, 88). We discovered two glycosyltransferase (GT2) enzymes (Fig. 4 and SI Appendix, Table S40) in Porphyra that are closely related to plant cellulose synthase-like (CSL) CSLAs (mannan synthases) and CSLs (xyloglucan synthases) and the green algal CSLA proposed to be implicated in mannan synthesis (89, 90). These Porphyra GT2 enzymes represent excellent candidates for mannan synthases, and their discovery in Porphyra suggests that both CSLA and mannan biosynthesis originated in the last common ancestor of green and red algae, rather than in the Viridiplantae, as previously proposed (89). These findings suggest that mannan had an ancestral function in red algae, although it appears the Florideophyceae abandoned mannan for cellulose, whereas the Bangiophyceae retain a mannan/synd-based cell wall in the gametophytes and a cellulose-based cell wall in the filamentous sporophytes. The Porphyra genome has two genes encoding glycosyl hydrolases (GH) of the GH13 family,
which are predicted to be β-mannanases, indicating that *Porphyra*
has the enzymes required for both the biosynthesis and degradation
of mannans present in the cell walls of the blade.

The extracellular matrix (ECM) component of the cell wall in
*Porphyra* is a highly hydrated agar called porphyrin that limits
desiccation and contributes to the flexibility of the cell wall
during osmotic stress (91) (*SI Appendix*, Figs. S41 and S45 and
Table S40). Porphyrin is mainly composed of porphyranobiose
but also contains varying percentages of agarobiose (92). The
*Porphyra* genome encodes a family 2 carbohydrate sulfotransferase
(ST2) that may act as a 1-galactose-6-O-sulfotransferase in
the biosynthesis of porphyrin (*SI Appendix*, Table S40). After
*C. crispus*, the closest related homologs to this red algal sulfotransferase are found in *metazoans* (*Nematostella vectensis*, sea
anemones) and the closest characterized enzymes are dermatan and chondroitin 4-O-sulfotransferases; similarly, the *Porphyra* GT7, after the *C. crispus* GT7, is most similar to *metazoan* GT7s (*Acronia digitifera*, staghorn corals) and the closest characterized enzyme is chondroitin sulfate N-acetylgalactosaminyl
transferase 2 (*Homo sapiens*). These findings lend support to the
ancestral nature of the biosynthesis of sulfated polysaccharides in
the eukaryotes. Homologous carbohydrate sulfotransferases are
indeed conserved at least in *metazoans*, brown algae, and red
algal, so were most likely present in the last-common eukaryotic
ancestor (91, 93). Genes encoding carbohydrate sulfotransferases
are not found in sequenced genomes of terrestrial plants and
freshwater algae, but marine angiopeptide have, relatively recently,
readapted to the marine environment by developing an ECM
containing sulfated polysaccharides through an unknown mecha-
nism of convergent evolution (94). This speaks to the critical role
of sulfated ECMs in high-salinity environments (physiological sa-
line and seawater). The *Porphyra* genome encodes three GH16
enzymes that form a clade with other red algae and with the
agarases and porphyranases from marine bacteria (*SI Appendix*,
Fig. S41). These enzymes likely function in cell wall remodeling
and are unknown in unicellular red algae, which supports the
possible involvement of bacteria in the evolution of multicellular
algae (95).

Obtaining Nutrients at High Tide. We found that *Porphyra* has genes
for Fe-uptake mechanisms that likely lend specificity and enable
high-affinity assimilation during the narrow tidal window when
blades are underwater (Fig. 5 and *SI Appendix*, Table S46). In
addition to several putative Fe$^{2+}$ transporters of the NRAMP
and ZIP family proteins (which are typically broad-specificity
divalent metal transporters) (96), the *Porphyra* genome en-
codes a high-affinity iron transport complex containing a per-
mease (FTR1) and multicopper oxidase (FET3 in yeast/FNX1 in
green algae) and members of the FEA and ISIP2A families.
The latter are related algal-specific protein families containing
secreted soluble proteins involved in iron assimilation and
membrane-bound iron-uptake facilitators (97, 98). Although
both iron-uptake strategies are found in microalgae, we were
surprised to find FTR1 and FOX1 homologs in a macroalgae,
given the absence of this complex in higher plants and animals.
Based on an analysis of phylogenetic distribution, FTR1 homologs
are found in each of the three main Archaeplastida lineages
(Viridiplantae, Glaucophyta, and Rhodophyta) but were lost after

Fig. 4. Neighbor-joining phylogenetic tree of GT2 showing the positions of *Porphyra* sequences (dark triangles) and rooted on bacterial cellulose synthase (bcsA). Red algal cellulose synthase (CESA) cluster most closely with streptomycete oomycete CESA, whereas bagniophyle CSL enzymes are sister to a clade of green plant CSLs that include a mannans synthase. Two red algal GT2s are related to lcsA, likely because of horizontal gene transfer. Red algal sequences lie on red branches and triangle group sequences in Viridiplantae (green), streptomycetes (brown), Cyanobacteria (turquoise), Bacteria (pink), and Alphaproteobacteria (gray). The proteins were aligned using MAFFT with the LINS-i algorithm and the scoring matrix Blosu62 and then manually refined with BUCKiT. The phylogenetic tree was calculated using maximum likelihood in MEGA 6.06. See *SI Appendix*, Table S42 for full names and GenBank accession numbers. Bootstrap values ≥65% are shown at nodes.

Bradley et al.

PNAS | Published online July 17, 2017 | 18067
the transition to land when plants evolved true roots (Fig. 5). Other inorganic and organic nutrient uptake and conversion capabilities appear similar to those described for other red algae (e.g., Fig. 5 and SI Appendix, Figs. S51 and Tables S49, S52, and S53), except that the number of ammonium transporters (seven genes) (SI Appendix, Table S49) and intracellular copper transporters (P-type copper transporters) (MTA1-60) (SI Appendix, Figs. S47 and Table S40) is greater in _Porphyra_, perhaps reflecting the metabolic demands of life in the upper intertidal zone.

**Methods**

_P. umbilicalis_ was isolated as an unripe culture from a blade growing at Schoodic Point, Maine (44°0.48' N, 68°32.14' W) on April 5, 2008. This single isolate (103) was cloned by culture of progeny from its sexual neutral, spores followed by DNA extraction, and assembly of the genome from sequences produced using Illumina and PacBio sequencing platforms (see SI Appendix, Methods for full details). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GI under accession no. MGAK00000000. The version described in this paper is version MGAK01000000.

ACKNOWLEDGMENTS. We are grateful to the anonymous reviewers and editor for their insightful and constructive suggestions. We thank Dr. Liibeth Miranda, Charlotte C. T. Quigley, and Charlotte Royer University of Maine for their major assistance in maintaining cultures of the genomic strain during the period project; Drs. Joyce Longcore (University of Maine) and M. Blackwell (Sullivans State University) for discussions about fungi; Dr. Sarah Topper Dronron for discussion about kelp slime elements; and the many colleagues who provided useful discussion as participants in the _Porphyra_ Research Coordination Network (supported by NSF OCE 2719472 to S.A., A.R.G., and J.W.S.), especially to Elisabeth Gartner (University of Maryland), an oral co-PI of the NSF RSN and DOE JGI contract. The work conducted by the US Department of Energy (DOE) Joint Genome Institute (JGI) had a DOE Office of Science User Facility, was supported by the Office of Science of the US DOE under Contract DE-AC02-05CH11231 (S.A., A.R.G., and J.W.S.). Other major research support provided was by NSF 0929598 to S.A. and A.R.G., National Oceanic and Atmospheric Administration (NOAA) Contract NAW16SAK70119 to S.A., DOE Office of Science, Office of Biological and Environmental Research (BER), and DOE Office of Energy Efficiency and Renewable Energy (EERE), and National Science Foundation (NSF).
7.4 Quantitative proteomics of a B12-dependent alga grown in co-culture with bacteria reveals metabolic trade-offs required for mutualism
Quantitative proteomics of a B_{12}-dependent alga grown in co-culture with bacteria reveals metabolic trade-offs required for mutualism

Katherine E. Hellwell1,2, Jagroop Pandhal2, Matthew B. Cooper1,3, Joseph Longworth2, Ulrich Johan Kudahl1, David A Russo2, Eleanor V. Tomsett1, Freddy Bunbury1, Deborah L. Salmon1, Nicholas Smirnoff3, Phillip C. Wright1,4, Alison G. Smith1

1 Department of Plant Sciences, University of Cambridge, Cambridge, CB2 3EA, UK
2 Department of Chemical and Biological Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK
3 Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, EX4 4QD, UK.

*These authors contributed equally to the work

Current addresses:
KEH – The Marine Biological Association of the UK, Plymouth, UK, PL1 2PB
MBC – Faculty of Biology, Medicine and Health, The University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK
DAMTR - Copenhagen Plant Science Centre, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871, Frederiksberg C, Denmark
PCW – Faculty of Science, Agriculture and Engineering, Newcastle University, Newcastle, NE1 7RU, UK

Corresponding authors:
Jagroop Pandhal, j.pandhal@sheffield.ac.uk, +44 (0)114 222 4914
Alison G Smith, as25@cam.ac.uk, +44 (0)1223 333952

Total word count for the main body of the text: 6470

Word counts for
Introduction: 1054 words
Materials and Methods: 1519
Results: 2293
Discussion: 1534
Acknowledgements: 70

Number of figures: 5 in total, Figs 1 & 2 in colour
Number of tables: 1
Supporting information: 3 Figures and 1 Table

Summary

- The unicellular green alga *Lobomonas rostrata* requires an external supply of vitamin B₁₂ (cobalamin) for growth, which it can obtain from the soil bacterium *Mesorhizobium loti* in exchange for fixed carbon in stable laboratory cultures. We investigated changes in protein expression that occur in the alga to allow it to engage in this mutualism.
- We developed a quantitative isobaric tagging (iTRAQ) proteomics platform to determine the *L. rostrata* proteome grown axenically with B₁₂ supplementation or in co-culture with *M. loti*. Data are available via ProteomeXchange (PXD005046).
- Using the related *Chlamydomonas reinhardtii* as a reference genome, 588 algal proteins could be identified. Enzymes of amino acid biosynthesis were higher in co-culture than axenic culture, and this was reflected in increased levels of total cellular protein and several free amino acids. A number of heat shock proteins were also elevated. Conversely, photosynthetic proteins and those of chloroplast protein synthesis were significantly lower in *L. rostrata* cells in co-culture. These observations were confirmed by measurement of electron transfer rates in cells grown under the two conditions.
- The results indicate that, despite the stability of the mutualism, *L. rostrata* experiences stress in co-culture with *M. loti*, and must adjust its metabolism accordingly.

Keywords: iTRAQ proteomics, *Lobomonas rostrata*, *Mesorhizobium loti*, Mutualism, Photosynthesis, Vitamin B₁₂
Introduction

Microalgae are a polyphyletic set of photosynthetic eukaryotes found across the eukaryotic tree of life that are estimated to be responsible for ~50% of global CO₂ fixation (Falkowski, 1998). In both soil and aquatic (marine and freshwater) habitats, algae exist alongside a wide spectrum of other microbes, including bacteria, archaea, fungi and cyanobacteria. This has led to millions of years of coevolution between contemporaneous species. It is therefore not surprising that a broad range of interactions between different microbial players within the ecosystem have been observed. As well as harmful interactions (Fernandes et al., 2011), many examples of beneficial associations have been documented between algae and bacteria, including promotion of algal growth (Park et al., 2008), support of cellular differentiation (Matsumoto et al., 2005), provision of bactericidal/algaecidal protection (Hold et al., 2001; Geng et al., 2008; Thiel et al., 2010), stress tolerance (Xie et al., 2013), and hormonal stimulation (Amin et al., 1999; Amin et al., 2015). Underpinning many mutualistic interactions is nutrient exchange (Cooper & Smith, 2015), where algal photosynthetic is exchanged for micronutrients provided by bacteria, for example by facilitation of iron uptake (Amin et al., 2009), or provision of the vitamins thiamine (vitamin B₁) (Paerl et al., 2015) or cobalamin (vitamin B₁₂) (Croft et al., 2005; Wagner-Döbler et al., 2010; Kazamia et al., 2012; Durham et al., 2015). These organic micronutrients are essential enzyme cofactors, but increasingly it is recognised that many algae, despite their photosynthetic lifestyles, require an exogenous source of one or more of these compounds, or their precursors, for growth (Croft et al., 2006). Analysis of environmental samples has detected the co-occurrence of bacterial producers and algal requiers of cobalamin in a variety of marine environments (Koch et al., 2013; Bertrand et al., 2015), and fertilisation experiments have demonstrated that these compounds are limiting for algal productivity (Koch et al., 2012), providing evidence that algobacterial interactions are likely to be widespread and of considerable significance for global net primary production.

Cobalamin is required by over 50% of all microalgal species surveyed (Croft et al., 2005), but with no phylogenetic relationship between dependent and non-dependent species, implying that this trait has arisen multiple times during algal evolution. The requirement for cobalamin is as a cofactor for B₁₂-dependant methionine synthase (METH), which is involved in the synthesis of the amino acid methionine, as well as more generally in cellular C1 metabolism. Species that do not require vitamin B₁₂ have an alternative, B₁₂-independent, enzyme, METE (Helliwell et al., 2011), which performs the same function, albeit at lower efficiency (González et al., 1992; Croft et al., 2005; Helliwell et al., 2015). Several species of algae, such as Chlamydomonas reinhardtii and Phaeodactylum tricornutum, encode both isoforms of methionine synthase, and can use METE when B₁₂ is not available. The presence of an external supply of the vitamin allows these species to use METH, and at the same time METE gene expression is repressed, both in laboratory cultures (Helliwell et al., 2011; Bertrand et al.,...
2012), and environmental samples (Bertrand et al., 2015). Long-term exposure to B12 in the environment might therefore lead to loss of the METE gene, and indeed a metE mutant was generated by an experimental evolution approach after growth of C. reinhardtii cells in B12 for ~500 generations (Helliwell et al., 2015). It is conceivable therefore that interactions with B12-producing bacteria in the environment over evolutionary time might lead to the frequent loss of METE in diverse algal lineages (Kazmania et al., 2016), which would account for the widespread occurrence of B12 auxotrophy across the algal lineages.

A major challenge now is to dissect the specific molecular mechanisms that underpin the exchange of cobalamin, but this is difficult in the dilute conditions of aquatic environment. To study aspects such as signalling, regulation, transporter proteins, and dynamics over time, defined model systems are required. In our laboratory we developed a model laboratory partnership between the vitamin B12-requiring freshwater green alga Lobomonas rostrata, a close relative of C. reinhardtii, and the rhizobial bacterium Mesorhizobium loti, which supplies vitamin B12 in exchange for fixed carbon (Kazmania et al., 2012). Although a synthetic interaction, the ease with which it formed mirrors that of other artificial systems such as the mutualism established between C. reinhardtii and the yeast Saccharomyces cerevisiae based on exchange of carbon and nitrogen (Horn and Murray, 2014). Moreover, physiological experiments with our L. rostrata and M. loti co-cultures under both batch and semi-continuous conditions revealed that an equilibrium in terms of cell number was established and maintained (Kazmania et al., 2012), demonstrating that the partnership exhibits a degree of regulation. The direct exchange of B12 could be modelled using growth dynamics of algal-bacterial co-cultures (Grant et al., 2014), and the model implicated specific release from live bacterial cells, rather than simply release of the vitamin once the cells had died and lysed. With this observation of true mutualism between the algae and bacteria, we might expect specific metabolic changes in L. rostrata when grown in co-culture compared to axenically with B12 supplementation.

Genome-scale approaches to dissect metabolic shifts occurring on account of interspecies microbial interactions have been applied to a variety of different phytoplankton-bacteria co-culture systems. Such work has revealed important insights into alterations in core metabolism, including amino acid biosynthesis in associations between cyanobacteria and heterotrophic bacteria (Beliaev et al., 2014; Biller et al., 2016), as well as sulphur cycling (Durham et al., 2015), and signalling and production of infochemicals (Amin et al., 2015) in diatom-bacterial interactions. However, the majority of the studies were at the transcript level, whereas it is proteins that are responsible for performing cellular processes. A requirement for proteomics analyses is the need for databases of peptides that can be matched to experimental mass spectra. Until recently this required annotated genome sequence information of the organism under study, which is not currently available for L. rostrata, but as more organisms have been sequenced, this has resulted in many more shared peptides in the databases. In
this study, we performed a series of preliminary bioinformatics and proteomics experiments to
evaluate the applicability of applying the global quantitative proteomics chemical isobaric tags
(iTRAQ) methodology with our system. Once the feasibility was confirmed, the approach was used to
identify metabolic differences in *L. rostrata* cells when grown in media supplemented with vitamin
B₁₂ compared to cultures where B₁₂ is provided by *M. loti* in a mutualistic exchange. This has provided
novel insight into the metabolic consequences of mutualism between a photosynthetic green alga and
its heterotrophic bacterial partner.

**Materials and Methods**

**Algal and bacterial strains and cultivation**

*Lobomonas rostrata* (SAG 45-1) was obtained from the Experimental Phycology and Culture
Collection of Algae at the University of Goettingen (EPSAG), Germany. It was grown autotrophically
on TP' medium (Kazamia et al., 2012). Vitamin B₁₂ was provided as cyanocobalamin (Sigma-Aldrich,
UK) at 100 ng/L, since this supports the maximum carrying capacity of *L. rostrata* (Kazamia et al.,
2012). The *L. rostrata*-*M. loti* co-culture was an established co-culture that had been maintained
growing over many generations without a source of organic carbon or vitamin B₁₂. Cultures were
maintained in 16h light: 8h dark cycle (16L:8D) with shaking (140 r.p.m.) at 25 °C. *M. loti* (MAFF
303099) was a gift from Prof Allan Downie, John Innes Centre, UK. It was maintained axenically in
TP+ with 0.1% v/v glycerol at 28 °C. Cells were harvested by centrifugation, and if not analysed
immediately, cell pellets were frozen in liquid N₂ and stored at -80 °C. Algal cell counts were
determined using a Dual Threshold Beckman Coulter (Z2) Particle Counter and Size Analyser.
Bacterial cell numbers were determined by plating on solid media.

**Protein preparation and iTRAQ labelling**

Cultures were harvested at 3000 g for 10 mins at 4 °C. The resulting cell pellets were further washed
in 0.5 M triethylammonium bicarbonate buffer (TEAB) prior to storage at -20 °C until ready for
analysis. Cells were defrosted and resuspended in 1 ml of 0.5 M TEAB incorporating a sonication bath
step for 5 minutes on ice. The resulting cell mix was ground in liquid nitrogen using a mortar and
pestle. The post grind sample was transferred to a low bind tube and the sonication bath treatment was
repeated, followed by two cycles of more intense sonication using a Micro tip Bronson Sonifier®
(Enerson, Danbury CT). Insoluble protein was removed by centrifugation at 18,000 g for 30 min at 4
C. The soluble protein fraction (supernatant) was quantified using RCDC manufacturers’ instructions (Bio Rad, Hercules, CA) and 100 µg protein was precipitated for 6 hours using acetone at -20 °C. The dried pellet was resuspended in 30 µl 500 mM TEAB (pH 8.5) with 0.1% sodium dodecyl sulphate and reduced with 2 µl 50 mM tris-(2-carboxyethyl)-phosphine at room temperature for 30 min, followed by alkylation with 1 µl 200 mM methyl methanethiosulfonate in isopropanol for 20 min in the dark. Samples were digested with 2.5 µg trypsin and 5 µl acetonitrile (ACN) at 37 °C for 16 h. The resulting peptides were either taken forward for mass spectrometry analysis for preliminary proteomics analysis or labelled with iTRAQ reagents following the manufacturer’s instructions (AB Sciex, Framingham, MA). Labels 113, 114, 115 and 116 were used to label biological replicates of *L. rostrata* cells cultured with B13 supplemented in the media, and labels 117, 118, 119 and 121 were used for *L. rostrata* cells cultured with *M. loti*.

**HILIC chromatography and mass spectrometry**

All 8 labelled peptide samples were combined before being dried in a vacuum concentrator (Concentrator 5301, Eppendorf, Stevenage(UK)). The sample was resuspended in 200 µl of buffer A (10 mM ammonium formate, 90% ACN, pH 3 [adjusted with formic acid]) and 100 µl was loaded onto a PolyHypereethyl A column (5 µm particle size, 20 cm length, 2.1 mm diameter, 200 Å pore size [PolyLC, Colombia, MD]) using an Agilent 1100-series HPLC (Agilent, Wokingham, U.K.). With a flow of 0.5 ml min⁻¹ buffer A was exchanged with buffer B (10 mM ammonium formate, 10% ACN, pH 4 [adjusted with formic acid]) to form a linear gradient as follows: 0% B (0-5 min), 0-15% B (5-7 min), 15% B (7-10 min), 15-60% B (10-50 min), 60-100% B (50-55 min), 100% B (55-65 min), 0%B (65-75 min). Beginning at 18 min, 22 fractions of 1 min length, followed by 3 fractions of 3 min length were collected and dried by vacuum centrifugation ready for reverse-phase liquid chromatography-tandem mass spectrometry (LC-MS/MS).

LC-MS/MS was conducted using an Ultimate 3000 HPLC (Dionex, Sunnyvale, CA, USA) coupled to a QStar XL Hybrid ESI Quadrupole time-of-flight tandem mass spectrometer (Applied Biosystems (now ABSciex), Framingham, MA, USA). Samples were resuspended in 20 µl buffer A (3% ACN, 0.1% FA) before loading 9 µl onto a Acclaim PepMap 100 C18 column, 3 µm particle size, 15 cm length, 75 µm diameter, 100 Å pore size (Dionex, Sunnyvale, CA, USA). With a flow of 300 µl min⁻¹ buffer A was exchanged with buffer B (97% ACN, 0.1% FA) to form a linear gradient as follows. 3% B (0 min), 3-35% B (5-95 min), 35-90% B (95-97 min), 90% B (97-102 min), 3% B (102-130 min). The mass detector range was set to 350-1800 m/z and operated in positive ion mode. Peptides with +2, +3, and +4 were selected for fragmentation. The mass spectrometry proteomics data have been
deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino et al., 2014) partner repository with the dataset identifier PXD005046.

Protein identification and quantification

The workflow for proteomic identifications is shown in Fig. S1. Searches were conducted using Mascot, Omssa, X!Tandem, Peaks and ProteinPilot against the Uniprot reference proteome for C. reinhardtii (Uniprot ID 3055) and M. loti (Uniprot ID 266835). Each search was conducted with the target-decoy database method to calculate false discovery rate (FDR). The decoy was formed using either reversed sequences (Mascot, Omssa, X!Tandem and ProteinPilot) or randomized sequence (Peaks) of the C. reinhardtii and M. loti proteomes. Searches were restricted to a peptide FDR of 3% prior to decoy hits being removed at the Peptide Spectral Match (PSMs) level. PSMs are the individual matches made by the search engine algorithm, between the mass spectrometers output product ion scan and the potential peptides contained within the search database. They combine to form the peptide and protein identifications. PSMs from the five search engines were merged using an R based script that also removed PSMs showing disagreement in terms of peptide assignment or protein identification between the search engines, as performed previously. Separately, reporter ion intensities for each PSM were extracted and matched to the merged results. For quantitation, these reporter ion intensities were extracted followed by variance stabilisation normalisation, isotopic correction and median correction, and finally averaging by protein. To avoid using arbitrary thresholds for highlighting differentially expressed proteins, a t-test was performed between replicate conditions to determine significance with fold change. Transmembrane domains were identified using TMHMM server 2.0 (http://www.cbs.dtu.dk/services/TMHMM/). We identified all putative transmembrane domains in each of the proteins identified in the iTRAQ experiment. If a sequence had 1 or more transmembrane domains this was classified as likely to be membrane-associated rather than soluble.

Analytical methods

Maximum potential photosynthetic capacity as electron transfer rate (ETR) was estimated using pulse amplitude modulation (PAM). Cells were prepared to a density of 2.5 \times 10^6 cells/ml, placed in cuvettes and mixed with a magnetic stirrer in order to maintain an even cell suspension. Instant light response curves of chlorophyll fluorescence were generated for each sample using a Mini-PAM (Heinz Walz GmbH, Effeltrich, Germany). The electron transfer rate (ETR) was calculated as follows:

\[ \text{ETR} = \frac{Y}{1000} \times \text{PAR} \times 0.5 \times \text{ETR factor} \]  

(1)
where $Y/1000$ is equal to $(Fm'-Fm)/Fm'$, or $\Phi_{PSII}$, the quantum yield of PSII. The ETR factor described refers to the fraction of incident photons absorbed by the sample, for which a default value of 0.83 was used (Maxwell & Johnson, 2000).

Protein was measured using Bradford’s reagent as described (Bradford, 1976) using bovine serum albumin as standard. For amino acid analysis, full experimental details are provided in the Supplementary Material. In brief, approximately 5 mg of a freeze-dried sample of cultures were extracted in acetonitrile, and then reextracted with 20% methanol spiked with an internal standard containing stable isotope-labelled amino acids (L-amino acid mix (Sigma-Aldrich, Co., St. Louis, MO, USA)). The supernatants were pooled and amino acids were derivatised with an AccQ-Tag Ultra derivatization kit (Waters Corporation, Milford, MA, USA). HPLC-ESI-MS/MS quantitative analysis of the amino acids was performed using an Agilent 6420B triple quadrupole (QQQ) mass spectrometer (Agilent Technologies, Palo Alto, USA). All ions were scanned in positive ion mode and given a dwell time of 50 msec. Data analysis was undertaken using Agilent Mass Hunter Quantitative analysis software for QQQ (Version B.07.01). Accurate quantification used the stable isotope labelled internal standards added during sample extraction, and data was normalised to the dry weight of the samples.

RT-qPCR

RNA was extracted from the cell pellet from 10 ml of day 14 (late-log) cultures using RNeasy® Plant Mini Kit (Qiagen), treated with Turbo DNA-free™ kit (Ambion) to remove genomic DNA, and reverse transcribed with Superscript III® First-strand synthesis system for RT-PCR (Invitrogen). Selected C. reinhardtii coding sequences from the NCBI database were aligned with scaffolds produced from L. rostrata RNA-seq data (unpublished) in order to find genes of interest; homologues were found to be 90% similar at the nucleotide level. Primers were designed with melting temperatures of 61-63°C, and to amplify products of 150-200 nucleotides (Untergasser et al., 2012), and the sequences are shown in Table S1. Quantitative PCR reactions were prepared in 10 µl volumes using SYBR® Green JumpStart™ Taq ReadyMix™ from (Sigma-Aldrich, UK). Reactions were performed in a Rotor-Gene Q (QIAGEN) with an initial denaturation at 95°C for 2 minutes followed by cycling 45 times between 58°C for 30 seconds and 95°C for 15 seconds, and finally a melt curve was generated by increasing temperature from 55 to 95°C in 1°C steps held for 5 seconds. Amplification efficiency and cycle threshold values were calculated with the Rotor-Gene Q Software version 2.02. Quantification and normalization was carried out using the Pfaffl model (Pfaffl, 2001) using three reference genes: Ubiquitin (UBQ), Eukaryotic translation initiation factor 4A (EIF4A) and Receptor of activated protein kinase C1 (RACK1) (Mus et al., 2007).
Results

Preliminary bioinformatics and proteomic analyses

To ascertain whether it was feasible to use iTRAQ for global analysis of the proteome of unsequenced L. rostrata, a series of preliminary experiments was carried out. The first aim was to find a suitable reference proteome for matching experimental spectra sourced from L. rostrata. This process relies on shared peptides from sequenced organisms being present within existing protein databases. The closely related freshwater green alga C. reinhardtii (Uniprot ID 3055) was a prime candidate and also has the advantage of having a relatively well annotated genome (https://phytozome.jgi.doe.gov/pz/portal.html). A search against C. reinhardtii produced 1% peptide spectral matches (PSM), similar to a previous cross-species proteomics study (Pandhal et al., 2008), implying that sufficient proteins could be identified in a global proteomics study to uncover biological insight into the metabolic processes of the unsequenced alga.

Our iTRAQ experiment was aimed at identifying proteins from organisms in co-culture, so proteins from both L. rostrata and M. loti would be present. Therefore it was important to consider the possibility that shared peptides between the two organisms might influence the results, namely that peptides from M. loti might be assumed to be from L. rostrata, and thus interfere with protein quantifications. To evaluate this, a theoretical tryptic digest was undertaken. The identification of shared peptides between the alga and bacteria was done using protein sequences for M. loti and C. reinhardtii retrieved from UniProt (02-07-2016). A python script was written to read the protein sequences and theoretically digest them into tryptic peptides, by cleaving them after arginine and lysine unless followed by a proline. Peptides between 6-16 amino acids long were identified and compared, calculating the shared fraction using the total number of unique peptides in C. reinhardtii. The result was that 0.31% of tryptic peptides within this size range were shared between C. reinhardtii and M. loti. Given this low number, and the close phylogenetic relationship between C. reinhardtii and L. rostrata, it was unlikely that interpretation of the iTRAQ experiment would be confounded by peptides shared between the alga and bacterium in the co-cultures.

Finally, the iTRAQ approach rests on the relative quantification of peptides across samples, with the general assumption that proteins are present in all samples, whereas in our study comparing co-cultures and axenic cultures of L. rostrata, only half would contain M. loti cells. Moreover, the aim of the experiment was to look for metabolic differences in L. rostrata under the two conditions, and therefore it would be ideal to minimise spectra sourced from M. loti, which could potentially reduce the total alga spectra generated per sample injection. This was potentially problematic as the ratio of
cell counts for L. rostrata to M. loti cells in stable co-cultures is between 1:30 and 1:100 (Kazamia et al., 2012), averaging 1:50 in the samples used for the iTRAQ experiment here. Moreover, proteins from sequenced organisms are more likely to be identified compared to those from unsequenced organisms, in bottom-up proteomics experiments. The latter was confirmed with proteins from axenic cultures of M. loti, where the PSM was 11% (compared to 1% PSM for L. rostrata). However, the much larger algal cell volume (~500 μm³) compared to ~0.5 μm³ for the bacteria means that in the co-culture this would correspond to significantly more algal protein compared to bacterial protein, even with the higher number of bacterial cells. In order to test this, a protein per cell measurement was made for axenic cultures of the two organisms, and found to be 2.29 x 10⁻⁷ mg cell⁻¹ for L. rostrata, and 3.25 x 10⁻⁹ mg cell⁻¹ for M. loti. At the cell densities in the co-culture this corresponds to approximately 0.45 mg ml⁻¹ for L. rostrata protein, compared to 0.033 mg ml⁻¹ for M. loti, a 14-fold higher amount. This, together with the nature of peptide fractionation in bottom-up mass spectrometry based proteomics, where there is a bias towards identification of high abundance peptides and thus proteins, provided confidence in our approach.

Overview of the L. rostrata proteome determined by iTRAQ analysis

Having established the parameters and methodology for the experiment, four biological replicates of both axenic L. rostrata supplemented with 100 ng/L B12, and established co-cultures of L. rostrata and M. loti, were grown in autotrophic medium. Cells were harvested after 14 days when the algae were at mid-late exponential growth phase (Fig. S2), and total soluble protein was extracted for proteomic analysis. Altogether, 588 proteins were identified based on 1 high confidence peptide hit, and about half of these (293) had a minimum of 2 high confidence peptide hits per protein. A total of 47 proteins were predicted to be membrane-associated by virtue of the presence of at least one TM domain. A search of the same mass spectral data against the M. loti database (Uniprot ID 266835) identified just four proteins (all based on 1 high confidence peptide hit), indicating that the vast majority of proteins identified were of algal origin. Comparison of proteins from the two different conditions revealed 153 proteins that were significantly differentially expressed (p<0.05). Of these, 70 were present in higher amounts in L. rostrata when grown with M. loti in comparison to the axenic condition, and 83 were less abundant. A heat-map representation of these data showing fold changes between co- and monoculture illustrates the consistency between individual replicates of the treatments (Fig. S3).

Comparative proteomics reveals fundamental changes in metabolism of L. rostrata when grown in co-culture with M. loti.
To establish how the metabolism of *L. rostrata* differed between the two treatments we used a combined KEGG/Mercator (Kanehisa & Goto, 2000; Lohse et al., 2014) based analysis to assign function to the differentially expressed proteins, and to classify them into particular cellular processes and metabolic pathways (Table 1, Fig. 1). This analysis revealed that the metabolic processes most affected were those involved in protein metabolism (38 proteins, 24.8% of total), amino acid metabolism (23 proteins, 15%) and photosynthesis (24 proteins, 15.7%). Within the first category, which included those involved general protein metabolism, translation, targeting and folding, about half were found in higher abundance in co-cultured *L. rostrata* compared to axenic culture, and half in lower amounts. Several chaperonins were elevated, along with cytosolic elongation and initiation factors. Four ribosomal proteins (S19, S20, L12 and L40) were in higher amounts, although another four (S2, S9, L3 & L8) were reduced. Strikingly, of the remaining 16 proteins of protein metabolism that were lower in cocultures, 15 were those of plastid protein synthesis, including several ribosomal proteins and elongation factors EF-Tu and EF-G.

This effect on chloroplast biology was further reflected in the substantial number of proteins associated with both the light-dependent reactions of photosynthesis and CO2 fixation that were much lower in the co-culture; none were elevated (Table 1). As seen in Fig. 2 (red boxes), there is reduced abundance of subunits of the chloroplast ATP synthase alongside those from the oxygen-evolving complex of photosystem II, cytochrome b6f complex (Rieske FeS protein, Re), and soluble electron carriers ferredoxin (Fd) and ferredoxin-NADP reductase (FNR). Several enzymes in the Calvin cycle were also in lower abundance, including RuBisCO activase, and two enzymes involved in photorespiration, hydroxypyruvate reductase and glycolate dehydrogenase (Chauvin et al., 2008). In addition, there was a decrease in two carbonic anhydrases, which catalyse conversion of carbon dioxide to bicarbonate, one of which showed the greatest log2-fold change (-2.91) of all detected proteins.

Subunits of soluble starch synthase (SS) and ADP-glucose pyrophosphorylase (STA1, STA6) both involved in starch synthesis were less abundant in the coculture, whereas starch phosphorylase (SP), important for starch remobilisation, was higher. Similarly, enzymes of glycolysis and the TCA cycle also showed a mixture of higher and lower abundance, but all the subunits of the mitochondrial electron transfer chain that were detected (three sub-units of NADH:ubiquinone oxidoreductase and two sub-units of mitochondrial ATP synthase) were reduced in co-cultures compared to axenic *L. rostrata*, as were enzymes involved in fatty acid biosynthesis, enoyl ACP reductase (EAR) and the biotin carboxylase subunit of the acetyl-CoA carboxylase (ACC) enzyme complex (Ohlrogge & Jaworski, 1997), and 3 out of 4 enzymes of tetrapyrrole biosynthesis, including two of the chlorophyll branch, light-dependent protochlorophyllide reductase and geranylgeranyl PP reductase. The
impression is that in L. rostrata grown with M. loti there is a reduction in biosynthesis of the
photosynthetic apparatus and activity.

In contrast, many enzymes of amino acid metabolism were in higher abundance in co-cultured L.
rostrata compared with monocultures, with a total of 18 enzymes involved in both the degradation and
biosynthesis of amino acids being elevated (Fig. 2). Amongst them were anthranilate synthase (ANS),
involved in the tryptophan biosynthesis pathway (Liu et al., 2010), cysteine synthase (O-acetylserine
sulphydrolase, OASTL4), the final enzyme of the synthesis of cysteine (Tai et al., 2001),
methylcrotonoyl-CoA carboxylase, used in the degradation of leucine to acetocacetate (Song et al.,
1994), and S-adenosylhomocysteine hydrolase (SAHH), involved in the methylation cycle of C1
metabolism (Palmer & Abeles, 1979). Just two proteins involved in amino acid metabolism were
lower in co-cultured L. rostrata, acetylglutamate kinase and argininosuccinate lyase, enzymes of the
urea cycle and important for production of fumarate for the TCA cycle (Shargool et al., 1988).
Ferredoxin-glutamate synthase (Fd-GOGAT), involved in nitrate assimilation, was also elevated in co-
cultures, as were two enzymes of S-assimilation. Two other enzymes of biosynthetic pathways were
elevated: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, an enzyme of the non-mevalonate
pathway for the production of the isoprenoid precursor isopentenyl-pyrophosphate (Eisenreich et al.,
2001), and THIC encoding an enzyme of de novo thiamine biosynthesis.

Of particular note in the other functional categories was the higher abundance in the co-cultures of
stress-response related proteins, comprising six proteins annotated as heat-shock proteins (HSPs): four
annotated as members of the HSP70 family, the others as a chloroplast-targeted Hsp101 homologue
and a HSP90-like protein. Also elevated were four flagellar-associated proteins and autophagy protein
(APG8), the latter being the most highly altered relative to axenic L. rostrata (3.84). And whilst 5
histones are lower in abundance in coculture, histone methylase and 3 other enzymes of DNA
metabolism were found in higher amounts.

Given the observed reduction in levels of almost all the proteins of the Calvin cycle, we decided to
determine if this was due to changes in transcript expression levels. We therefore carried out RT-
qPCR using as template RNA extracted from cultures grown in conditions identical to those used for
proteomics analysis. We were able to identify sequences for eight proteins involved in carbon fixation
by searching an RNA-seq dataset from L. rostrata (U.J.H. K.E.H and A.G.S, unpublished) using the C.
reinhardtii homologues, to allow design of appropriate PCR primers. We also found L. rostrata
homologues for three of the most elevated proteins, THIC, APG8, and GAP1A encoding cytosolic
glyceraldehyde 3-phosphate dehydrogenase. Fig. 3 shows the results in order of the fold-abundance
detected in the iTRAQ experiment, with those lower in co-culture than in monoculture on the left, and
those higher on the right. In general, the trend in the transcript levels mirrored that of the proteins,
with significant (p<0.001) up-regulation of GAP1A and APG8 in co-culture (roughly 3-fold and 5-fold respectively). The thiamine biosynthesis enzyme THIC was also elevated, although this was not statistically significant. Similarly, there was significant (p-value <0.001) down-regulation of phosphoglycerate kinase 1 (PGK1) and chloroplast glyceraldehyde 3-phosphate dehydrogenase 3 (GAP3), as well as sedoheptulose-1,7-bisphosphatase (SERP1; p<0.01). The transcript levels for the other photosynthetic genes were unaltered or slightly higher in co-culture but not significantly. It should be mentioned that the CAHI gene, encoding a carbonic anhydrase, may encode a different isoform to the proteins detected in the iTRAQ experiment.

Photosynthetic electron transfer rate (ETR) is reduced in co-culture with M. loti

Our observation that several proteins involved in both the light reactions of photosynthesis and carbon dioxide fixation were less abundant in L. rostrata cells in co-culture with M. loti prompted us to investigate whether these altered protein levels impacted photosynthetic activities of cells in the symbiotic versus axenic treatment. Accordingly, we made photo-physiological measurements of cells in the two treatments using pulse amplitude-modulated fluorimetry (PAM). Quantification of the quantum yield of photosystem II and photosynthetic electron transfer rate (ETR) provides information concerning the overall photosynthetic performance of a culture. L. rostrata cultures, either with 100 ng/L B12 or with M. loti, were grown in identical conditions to those described for the proteomics experiment. Measurements of ETR were made at 7, 11 and 14 days after inoculation into fresh medium, corresponding to three different growth stages of the culture (Fig. S2). As can be seen in Fig. 4, at 7 days, there is little difference in ETR, but at both days 11 and 14 it was significantly lower in cells grown in the presence of M. loti, with the effect becoming more pronounced, so that at day 14 (corresponding to samples used for the proteomics analysis) the rate was only about 60% of that in axenic cultures.

L. rostrata cells grown in co-culture with M. loti have higher levels of amino acids and total protein

The other major group of proteins that were altered were enzymes of amino acid biosynthesis. Accordingly we measured total protein in the cultures and found that for both the cell pellet and the media it was approximately double in the co-culture compared to the axenic cultures (Fig. 5a). Whilst the former contain bacterial cells in addition, the estimated contribution of the bacterial cells to the total protein is likely less than 10% (see above). We also determined the profile of free amino acids in the cells using HPLC-mass spectrometry (Fig. 5b, note the log scale for the y-axis). The majority of the 15 protein amino acids that could be identified by this method were elevated in co-culture, with up to 2-fold changes seen for asparagine (Asn), glutamate (Glu) and histidine (His). The exceptions were
methionine and serine, where slight decreases were observed. Given that methionine is the product of the enzyme in *L. rostrata* that requires B_{12} (cobalamin-dependent methionine synthase, METH), this might indicate B_{12} limitation, a conclusion supported by elevation of SAHH in co-culture, and the observation that addition of B_{12} to the media of co-cultures enhances growth of *L. rostrata* (Kazarnia *et al.*, 2012).

**Discussion**

In this study we have applied quantitative iTRAQ proteomics to understand the metabolic differences in *L. rostrata* cells when grown in media supplemented with vitamin B_{12} compared to cultures where B_{12} is provided by *M. loti* cells. *L. rostrata* does not have a sequenced genome, and therefore its proteome is unknown, and the presence of bacterial cells within the co-culture samples also complicated the analysis. A series of preliminary experiments were undertaken, an approach that is essential for these sorts of co-culture, cross-species proteomics experiments. Through these analyses we demonstrated that (i) *C. reinhardtii* would provide a suitable reference database by a 1% PSM rate to *L. rostrata* peptides, (ii) that a low (0.31%) shared tryptic peptide (6-16 amino acids length) rate between *C. reinhardtii* and *M. loti* minimised interference by shared peptides, and (iii) the ~14 fold less bacterial protein compared to algal protein in the co-culture would not be detrimental to the overall algal protein numbers confidently identified and relatively quantified using iTRAQ. The value of these preliminary experiments was successfully demonstrated during the iTRAQ experiment, where we were able to identify 588 *L. rostrata* proteins with at least 1 high confidence peptide hit. Protein fold changes were comparable to previous iTRAQ studies, where relative values are significantly lower than usually seen for transcript level quantifications. To avoid using arbitrary thresholds for highlighting differentially expressed proteins, stringent quantitative statistics were applied (Noirel *et al.*, 2011) and revealed 153 proteins significantly differentially expressed (*p*-value < 0.05). (Table 1, Fig. 1). Furthermore, whilst transcript expression may not necessarily reflect protein levels, qRT-PCR analysis showed transcript and protein expression correlated in the two treatments for several of the genes we sampled (Fig. 3), providing further verification of our approach.

Inspection of the classes of protein that were altered revealed a higher abundance of proteins related to amino acid biosynthesis (Fig. 2a), and this was accompanied by modest but significant increases in several protein amino acids in the algal cells (Fig. 5b). *M. loti* is dependent on *L. rostrata* for fixed carbon in the co-cultures, but the specific compound, or compounds, that the alga supplies is currently unknown. It is possible therefore that amino acids are involved, as has been documented for several
other symbiotic interactions. For example, growth of the heterotrophic bacterium Shewanella W3-18-1
with the cyanobacterium Synechococcus sp. PCC 7002 modulates amino acid metabolism in
Shewanella (Beliaev et al., 2014). A reduction in expression of genes involved in alanine and
methionine biosynthesis in this bacterium has resulted from enhanced excretion of these
amino acids by the cyanobacterial partner, indicative of metabolite exchange. Perhaps the best-known
example of amino acids is provided by legume-rhizobial interactions, where the nitrogen fixed by the
symbiotic bacteria is transferred to their legume partners in the form of ammonium, some of which is
metabolised by the plant and returned to the bacteria in the form of amino acids in order to support
their growth (Lodwig et al., 2003). There are further similarities between proteins significantly up-
regulated in co-cultured L. rostrata and those in the nodule symbiosome, the membrane bordering the
nitrogen-fixing root nodule. These include a 60 kDa chaperonin, a homologue of that found in
symbiosomes of Glycine max (Panter et al., 2000), heat shock protein 70 found in symbiosomes of
Lotus japonicus and Medicago truncatula (Wienkopp & Saalbach, 2003; Catalano et al., 2004), and
calreticulin also found in L. japonicus (Wienkopp & Saalbach, 2003). The canonical function of
chaperonins and heat-shock proteins is in assisting correct protein folding, and prevention of protein
aggregation in cells undergoing some form of stress (Saibil, 2013). Other proteins up-regulated in co-
cultured L. rostrata that are also found in the symbiosome membrane are D-3-phosphoglycerate
dehydrogenase involved in amino acid biosynthesis (Wienkopp & Saalbach, 2003), citrate synthase
(Saalbach et al., 2002), cysteine synthase (Wienkopp & Saalbach, 2003) and lysine-tRNA ligase
(Catalano et al., 2004). The prokaryotic isoforms of some of these proteins have also been found to be
present in the symbiosome space (the area between the plant and bacterial cell membranes in the
symbiosome) rather than the membrane, indicating that they share a common symbiotic function
between kingdoms (Emerich & Krishnan, 2014). The exact function these proteins have acquired in
the symbiosome is still mostly unknown. In other symbiotic systems, mostly parasitic, the prokaryotic
isoforms of these proteins are thought to be important in the cell-cell adhesion of symbiont and host
and it could be that the same is true for the eu-karyotic isoforms (Copley, 2012), although it is known
that cell-to-cell contact is not an obligate requirement for our L. rostrata and M. loti system (Kazmania
et al., 2012).

Perhaps the most striking feature of L. rostrata grown in co-culture with M. loti is that there is overall
reduction in many chloroplast proteins. Firstly, proteins involved in chloroplast protein synthesis and
import are much lower in co-culture, whereas several of the cytosolic equivalents are elevated.
Correlating with this observation is a reduction in two chloroplast-encoded subunits of the ATP
synthase. At the same time, many components of the nucleus-encoded light-dependent reactions of
photosynthesis (Fig. 2), including the Rieske Fe-S subunit (Re) of the cytochrome complex, ferredoxin
(Fd), and Fd-NADP reductase (FNR), and virtually all of the Calvin cycle enzymes are also found in
lower amounts in the co-culture cells. Measurement of the ETR confirmed that photosynthetic
capacity of *L. rostrata* cells grown with *M. loti* in co-culture is much less than in axenic culture (Fig. 4). This is in contrast to a transcriptomics analysis of the cyanobacterium *Prochlorococcus* NATL2A
grown in co-culture with the heterotrophic bacterium *Alteromonas macleodii* MIT 1002, where a
notable enhancement of genes encoding photosynthetic apparatus, including subunits of PSI and
chlorophyll biosynthesis proteins, was seen in the cyanobacterium. The profound impact interactions
can have on growth and chlorophyll fluorescence parameters has also been highlighted by a systematic
physiological study that screened a library of hundreds of heterotrophic marine bacteria in pairwise
co-cultures with different *Prochlorococcus* ecotypes (Sher et al., 2011). In general, this study
identified that ‘enhancing interactions’, whereby the presence of heterotrophic bacteria enhanced
cyanobacterial growth rate and/or chlorophyll fluorescence, were more common than the other way
round. However, in this work the treatment conditions were such that all the necessary nutrients for
growth were present in the medium (in this case Pro99 medium supplemented with pyruvate, acetate,
lactic, glycerol, alongside vitamins) even in the presence of the bacterium, which is in contrast to our
set-up in which the co-culture treatment lacked a bacterial carbon source, alongside a vital
micronutrient necessary for *L. rostrata* growth. Thus in a scenario where growth of both microbial
partners is absolutely dependent of the presence of the other, a fine balance between mutualism and
competition appears to be at play. Whilst *L. rostrata* growth (and photosynthetic output) is supported
by a B12-synthesising bacterium, it is nevertheless constrained compared to growth in media where all
necessary nutrients are in adequate supply. There may also be a degree of competition between the
two species for other nutrients in the media (such as nitrogen and trace elements) and the co-cultured
algae may place its efforts into producing resource acquisition proteins rather than photosynthesis
related proteins. In fact in the closely related algae *C. reinhardtii* it is well known that photosynthetic
gene expression declines during early nutrient deprivation (Zhang et al., 2004; Schmolzinger et al.,
2014), which has been suggested to be due in part to accumulation of transcripts for *rhlA*, a putative
transcriptional repressor of nuclear-encoded chloroplast proteins (Nedelcu, 2009). This competition
may also provide another explanation for the up-regulation of amino acid biosynthesis enzymes in co-
cultured *L. rostrata*, the co-cultured algae requiring more resource acquisition machinery rather than
providing amino acids to its symbiont (Arrigo, 2005). The

A previous study of a mutualistic co-culture based on B12-exchange, in this case between the marine
diatom *Thalassiosira pseudonana* and a roseobacter *Ruegeria pomeroyi* DSS-3, looked at transcript
level changes, and identified that the likely compound supplied by the alga was the sulphur-containing
compound 2,3-dihydroxypropane-1-sulphonate (Durham et al., 2014). They found no evidence of
alterations in transcripts for enzymes of the photosynthetic machinery, nor any change in
photosynthetic activity. In contrast, the results of our study at the protein level, which is a more direct
indication of cellular activity, suggest – perhaps counter-intuitively – that a photosynthetic organism
grown mutualistically with a bacterium does not increase its photosynthetic activity to support both
organisms. Instead, the photosynthetic capacity of the alga is likely constrained more by its own nutrient status than it is by the demands of its partner. In this system, where nutrient (B12) status of the alga is effectively set by the bacterium, and vice versa (in terms of carbon supply), there appears to be a trade-off in terms of the algal response. The striking impact co-habiting bacteria can have on algal physiology, evolution and metabolism, is becoming increasingly recognised. Algae and bacteria are amongst the most abundant organisms on the planet and together their metabolism impacts global nutrient cycling and energy flow. Thus enhancing our knowledge of the molecular basis of algal-bacterial interactions is critical to establish the fundamental core principles governing microbial community function in the aquatic biosphere.

Acknowledgements

The work in this manuscript was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK (grant BB/I013164/1, a CASE studentship for M.B.C joint with PML Applications Ltd, Plymouth, UK, and a DTP studentship for F.B.); a EU FP7 Marie Curie ITN PhotoComm. no. 317184 for U.J.K; Engineering and Physical Science Research Council (EPSRC) grant EP/E036252/1 and the Natural Environment Research Council (NERC) grant NE/J024767/1.

Author Contributions

J.P., K.E.H., P.C.W. and A.G.S conceived the project and designed the research; K.E.H., M.B.C., J.L., U.J.K, D.A.R., E.V.T., F.B., D.S. and N.S. carried out the research; K.E.H., M.B.C, U.J.K, F.B., N.S., J.P. and A.G.S. carried out the data analysis and interpretation; K.E.H., M.B.C., J.P. and A.G.S. wrote the manuscript, with input from all authors. All authors read and approved the final manuscript.

References


in plant science 6: 78–84.
20

Frontiers in Microbiology 3: 363.


215

Proteomics 8: 2266–2284.


Nucleic acids research 29: e45.


Figure Legends

Figure 1 Overview of annotated functions ascribed to L. rostrata proteins. Proteins were extracted from 14 day cultures of L. rostrata grown in co-culture with M. loti, or axenically supplemented with 100 ng/L cyanocobalamin (vitamin B12). (a) Relative proportions of each protein category determined by the iTRAQ experiment to be more abundant in co-culture (a total of 70 proteins); (b) Relative proportions of each protein category found in lower amounts (a total of 83 proteins). Proteins were categorised into major functional groups based on combined KEGG/Mercator (Kanehisa & Goto, 2000; Lohse et al., 2014) analysis.

Figure 2 Schematic diagram summarising changes in primary metabolic enzymes of L. rostrata in co-culture versus in axenic culture with B12 supplementation. Proteins found in higher amounts in co-culture are indicated in blue, whereas those less abundant are shown in red. Abbreviations are provided in Table 1.

Figure 3 RT-qPCR measurements of transcript levels. Transcripts of genes encoding enzymes of photosynthesis that were found in lower abundance in co-cultures, plus three proteins highly elevated in co-cultures were quantified using cDNA prepared from RNA extracted from 14 day-old cultures. The transcripts labelled on the x-axis are in ascending order of protein fold-change measured by the iTRAQ experiment. The vertical dashed line separates proteins lower in abundance in co-culture (left), from those higher (right). Transcript abundance was normalised against 3 housekeeping genes (EIF4A, UBQ, RACK1) and then levels in co-cultures with M. loti (black columns) shown relative to...
the level in axenic _L. rostrata_ cultures (grey columns) set as 1. Error bars = SE, n=3, p-value <0.01=*, <0.001=** respectively (Students t-test). Abbreviations are as in Table 1.

**Figure 4 Photosynthetic capacity of _L. rostrata_ cultures.** PAM measurements of the electron transfer rate (ETR) were made to estimate overall photosynthetic capacity. Solid line shows data from cultures grown axenically with 100 ng/l cyanocobalamin, while dashed lines shows ETR from co-cultures with _M. loti_. Rates were measured in samples taken from cultures at days 7, 11 and 14, diluted to equivalent cell densities (2.5 × 10⁶ cells/ml). Data are the mean of 3 replicates ± SEM.

**Figure 5 Protein and amino acid content of _L. rostrata_ cultures.** (a). Total protein from the cell pellet or media of 14 day old cultures of axenic _L. rostrata_ or in co-culture with _M. loti_. Protein is expressed in picograms per algal cell relative to a standard of bovine serum albumin. (b). Amino acids from the cell pellet on day 14 were separated by HPLC and quantified by mass spectrometry. Amino acids are labelled on the x axis in alphabetical order, and quantities are expressed in femtograms per algal cell on the y axis. Axenic _L. rostrata_ + 100 ng/l B₁₂ = grey columns, _L. rostrata_ + _M. loti_ = black columns. Error bars = SE, n=4, p-value<0.05=*, <0.01=**. (Students t-test).
Table 1 List of identified proteins altered in abundance in co-cultures of L. rostrata versus reinhardtii protocone (indicated by the Uniprot ID). A total of 70 proteins were found to be significantly ($p < 0.05$) more abundant in the cocultures with M. loti (positive values), and 83 were less abundant (negative values). Identified proteins are grouped into functional categories as in Figure 1, ranked according to fold change. Data are available via ProteomeXchange (PXD005046). *also involved in TCA cycle

<table>
<thead>
<tr>
<th>General Function</th>
<th>Uniprot entry</th>
<th>Protein ID</th>
<th>Protein function</th>
<th>Fold-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-3</td>
<td>ABISZ1</td>
<td>eEF1a1</td>
<td>Eukaryotic translation elongation factor 1 alpha 1 (EC 3.6.5.3)</td>
<td>2.69</td>
</tr>
<tr>
<td>L112</td>
<td>ABISR7</td>
<td>RPL12</td>
<td>Ribosomal protein L112</td>
<td>2.42</td>
</tr>
<tr>
<td>60</td>
<td>I2FKQ9</td>
<td>CPN60C</td>
<td>Mitochondrial chaperonin 60</td>
<td>2.41</td>
</tr>
<tr>
<td>520</td>
<td>ABJ8M9</td>
<td>RPS20</td>
<td>Ribosomal protein S20</td>
<td>2.18</td>
</tr>
<tr>
<td>1.6</td>
<td>ABJ7C8</td>
<td>KIR5</td>
<td>Lysine–tRNA ligase (EC 6.1.1.6) (Lysyl-tRNA synthetase)</td>
<td>2.00</td>
</tr>
<tr>
<td>1</td>
<td>Q66YD0</td>
<td>VPPP1</td>
<td>Chloroplast vesicle-inducing protein in plastids 1</td>
<td>1.98</td>
</tr>
<tr>
<td>50</td>
<td>ABJ5Z0</td>
<td>RPP0</td>
<td>Acidic ribosomal protein P0</td>
<td>1.93</td>
</tr>
<tr>
<td>19</td>
<td>AB403</td>
<td>RPS19</td>
<td>Ribosomal protein S19</td>
<td>1.89</td>
</tr>
<tr>
<td>72</td>
<td>ABHTK7</td>
<td>TBA1</td>
<td>PabA translation factor</td>
<td>1.72</td>
</tr>
<tr>
<td>60A</td>
<td>ABII7B</td>
<td>CPN60A</td>
<td>Chaperonin 60A</td>
<td>1.72</td>
</tr>
<tr>
<td>1.6</td>
<td>P25840</td>
<td>HSF70</td>
<td>Heat shock 70 kDa protein</td>
<td>1.70</td>
</tr>
<tr>
<td>60B2</td>
<td>ABITH8</td>
<td>CPN60B2</td>
<td>Chaperonin 60B2</td>
<td>1.52</td>
</tr>
<tr>
<td>3.6.5.3</td>
<td>AB10X9</td>
<td>EFG2</td>
<td>Elongation factor 2 (EC 3.6.5.3)</td>
<td>1.51</td>
</tr>
<tr>
<td>40</td>
<td>ABICX9</td>
<td>RPL40</td>
<td>Ribosomal protein L40</td>
<td>1.44</td>
</tr>
<tr>
<td>40</td>
<td>AB10Q1</td>
<td>IP41</td>
<td>Importin subunit alpha</td>
<td>1.33</td>
</tr>
<tr>
<td>29</td>
<td>AB2322</td>
<td>EIF2G</td>
<td>Eukaryotic initiation factor</td>
<td>1.28</td>
</tr>
<tr>
<td>20</td>
<td>AB10E0</td>
<td>CPN20</td>
<td>Chaperonin 20</td>
<td>1.17</td>
</tr>
<tr>
<td>22</td>
<td>ABIVS9</td>
<td>TIC22</td>
<td>Elongation factor Tu at the inner membrane of chloroplasts (Fragment)</td>
<td>-1.14</td>
</tr>
<tr>
<td>Cysteine endopeptidase</td>
<td>ABJ7T7</td>
<td>CEP1</td>
<td>-1.16</td>
<td></td>
</tr>
<tr>
<td>Plastid ribosomal protein S1 (Ribosomal protein S1 homologue)</td>
<td>Q70DX8</td>
<td>S1PRPS1</td>
<td>-1.27</td>
<td></td>
</tr>
<tr>
<td>Elongation factor Tu, chloroplastic (EF-Tu)</td>
<td>P177A6</td>
<td>TUF</td>
<td>-1.30</td>
<td></td>
</tr>
<tr>
<td>605 ribosomal protein L3 (Fragment)</td>
<td>E35C57</td>
<td>RPL13</td>
<td>-1.34</td>
<td></td>
</tr>
<tr>
<td>Elongation factor Ts</td>
<td>Q5QE82</td>
<td>EFTs</td>
<td>-1.34</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein S2</td>
<td>ABHM64</td>
<td>RPS2</td>
<td>-1.37</td>
<td></td>
</tr>
<tr>
<td>Chloroplast elongation factor G (EC 3.6.5.3)</td>
<td>AB10A9</td>
<td>EFG1</td>
<td>-1.41</td>
<td></td>
</tr>
<tr>
<td>Peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8) plastid</td>
<td>AB1282</td>
<td>CYN20</td>
<td>-1.41</td>
<td></td>
</tr>
<tr>
<td>405 ribosomal protein S9 (Fragment)</td>
<td>E35C50</td>
<td>RPS9</td>
<td>-1.42</td>
<td></td>
</tr>
<tr>
<td>41 kDa ribosome-associated protein (chloroplast)</td>
<td>Q66Y83</td>
<td>RAP41</td>
<td>-1.49</td>
<td></td>
</tr>
<tr>
<td>38 kDa ribosome-associated protein (Chloroplast stem-loop-binding protein)</td>
<td>Q66Y82</td>
<td>RAP38</td>
<td>-1.52</td>
<td></td>
</tr>
<tr>
<td>505 ribosomal protein L5, chloroplastic</td>
<td>Q8HTL1</td>
<td>RPL5</td>
<td>-1.52</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Enzyme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABVX1</td>
<td>RPL8 Ribosomal protein L8</td>
<td>-1.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q84J21</td>
<td>ERY2 50S ribosomal protein L22, chloroplastic</td>
<td>-1.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q75258</td>
<td>PRPS5 Ribosomal protein S5 (plastid)</td>
<td>-1.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q86G23</td>
<td>NAB1 Nucleic acid binding protein (Putative nucleic acid binding protein)</td>
<td>-1.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A8J824</td>
<td>PRPL1 Plastid ribosomal protein L1</td>
<td>-1.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A8J8U6</td>
<td>CYN38 Peptidyl-prolyl cis-trans isomerase, cyclophilin-type (plastid)</td>
<td>-1.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D20032</td>
<td>RPS18 30S ribosomal protein S18, chloroplastic</td>
<td>-1.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A8VWS8</td>
<td>PRPL28 Plastid ribosomal protein L28</td>
<td>-2.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Photosynthetic electron transfer**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>A8V40</td>
<td>Fd Apoferrodoxin</td>
<td>-1.18</td>
</tr>
<tr>
<td>A8999</td>
<td>Re Rieske ferredoxin</td>
<td>-1.19</td>
</tr>
<tr>
<td>A8J0E4</td>
<td>PSBO Oxygen-evolving enhancer protein 1 of photosystem II</td>
<td>-1.21</td>
</tr>
<tr>
<td>A8HXL8</td>
<td>ATPC Chloroplast ATP synthase gamma chain</td>
<td>-1.25</td>
</tr>
<tr>
<td>B7U1J0</td>
<td>atpA ATP synthase subunit alpha, chloroplastic (EC 3.6.3.14)</td>
<td>-1.32</td>
</tr>
<tr>
<td>Q8PS4</td>
<td>HYDG Fe-hydrogenase assembly protein (Hydrogenase assembly factor)</td>
<td>-1.36</td>
</tr>
<tr>
<td>A8J6Y8</td>
<td>FNR2 Ferredoxin–NADP reductase (EC 1.18.1.2)</td>
<td>-1.63</td>
</tr>
<tr>
<td>P53991</td>
<td>PETH Ferredoxin–NADP reductase, chloroplastic (FNR) (EC 1.18.1.2)</td>
<td>-1.66</td>
</tr>
</tbody>
</table>

**Carbon assimilation**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>A8VM9</td>
<td>GCS5 Glycine cleavage system, P protein (EC 1.4.4.2)</td>
<td>-1.15</td>
</tr>
<tr>
<td>Q6A05</td>
<td>RCA1 Rubisco activase</td>
<td>-1.23</td>
</tr>
<tr>
<td>P24489</td>
<td>RA Ribulose bisphosphate carboxylase/oxygenase activase, chloroplastic</td>
<td>-1.36</td>
</tr>
<tr>
<td>Q0Z21</td>
<td>GDH Glycolate dehydrogenase</td>
<td>-1.37</td>
</tr>
<tr>
<td>A8HP84</td>
<td>GAP3 Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) (EC 4.3.2.1)</td>
<td>-1.47</td>
</tr>
<tr>
<td>A8P17</td>
<td>HPB1 Hydroxybutyrate reductase</td>
<td>-1.51</td>
</tr>
<tr>
<td>A8R2K4</td>
<td>SEB1 Sedoheptulose-1,7-bisphosphatase (EC 3.1.3.37)</td>
<td>-1.53</td>
</tr>
<tr>
<td>A8IKW6</td>
<td>RPE1 Ribulose-phosphate 3-epimerase (EC 5.1.3.1)</td>
<td>-1.57</td>
</tr>
<tr>
<td>A85A0</td>
<td>CPLD45 Predicted protein</td>
<td>-1.58</td>
</tr>
<tr>
<td>A8JKQ</td>
<td>FBP1 Fructose-1,6-bisphosphatase (EC 3.1.3.11)</td>
<td>-1.65</td>
</tr>
<tr>
<td>Q42690</td>
<td>FBA1 Fructose-bisphosphate aldolase 1, chloroplastic (EC 4.1.2.13)</td>
<td>-1.71</td>
</tr>
<tr>
<td>A8IRQ1</td>
<td>RPI1 Ribose-5-phosphate isomerase (EC 5.3.1.6)</td>
<td>-1.72</td>
</tr>
<tr>
<td>Q75N22</td>
<td>LC8 Low-CO2 inducible protein LC8</td>
<td>-1.77</td>
</tr>
<tr>
<td>A8J0C4</td>
<td>PGK1 Phosphoglycerate kinase (EC 2.7.2.3)</td>
<td>-1.81</td>
</tr>
<tr>
<td>P93109</td>
<td>CA2 Beta-carbonic anhydrase</td>
<td>-1.95</td>
</tr>
<tr>
<td>Q35859</td>
<td>BETA-CA1 Carbonic anhydrase</td>
<td>-2.91</td>
</tr>
</tbody>
</table>

**Starch metabolism**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2VA40</td>
<td>PHOA Phosphorylase (EC 2.4.1.1)</td>
<td>1.48</td>
</tr>
<tr>
<td>Q2VA41</td>
<td>PHOB Phosphorylase (EC 2.4.1.1)</td>
<td>1.42</td>
</tr>
<tr>
<td>A8H514</td>
<td>STA1 ADP-glucose pyrophosphorylase (EC 2.7.7.27)</td>
<td>-1.23</td>
</tr>
<tr>
<td>Q86L16</td>
<td>STA6 ADP-glucose pyrophosphorylase (EC 2.7.7.27)</td>
<td>-1.34</td>
</tr>
<tr>
<td>O64926</td>
<td>SS Soluble starch synthase (EC 2.4.1.21)</td>
<td>-1.81</td>
</tr>
</tbody>
</table>

**Glycolysis**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>A8JHR9</td>
<td>GAP1A Glyceraldehyde-3-phosphate dehydrogenase, dominant splicing variant (EC 1.2.1.12)</td>
<td>3.01</td>
</tr>
<tr>
<td>A8IEZ3</td>
<td>PG1 Glucose-6-phosphate isomerase (EC 5.3.1.9)</td>
<td>-1.22</td>
</tr>
<tr>
<td>A8V6R6</td>
<td>PYK2 Pyruvate kinase (EC 2.7.1.40)</td>
<td>-1.70</td>
</tr>
<tr>
<td>Enzyme Code</td>
<td>Gene Symbol</td>
<td>Enzyme Name</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>A8I20</td>
<td>PGX1</td>
<td>Phosphoglycerate kinase (EC 2.7.2.3)</td>
</tr>
<tr>
<td>A8I250</td>
<td>CIS2</td>
<td>Citrate synthase (EC 2.3.3.1)</td>
</tr>
<tr>
<td>A8I306</td>
<td>CIS1</td>
<td>Citrate synthase</td>
</tr>
<tr>
<td>A8H97</td>
<td>ACH1</td>
<td>Aconitase hydratase (EC 4.2.1.3)</td>
</tr>
<tr>
<td>Q9F6S5</td>
<td>MDH5</td>
<td>NADP-Malate dehydrogenase (EC 1.1.1.82)</td>
</tr>
<tr>
<td>A8HPL8</td>
<td>DLD2</td>
<td>Dihydrodipalmityl dehydrogenase (EC 1.8.1.4)</td>
</tr>
<tr>
<td>A8J776</td>
<td>DLA2</td>
<td>Dihydrolipoyl dehydrogenase (EC 2.3.1.12)</td>
</tr>
<tr>
<td>A8JDV9</td>
<td>ATP3</td>
<td>F1F0 ATP synthase gamma subunit</td>
</tr>
<tr>
<td>Q9F6S5</td>
<td>ATPA</td>
<td>ATP synthase subunit alpha</td>
</tr>
<tr>
<td>Q9F6S5</td>
<td>NUO58</td>
<td>NADH:ubiquinone oxidoreductase subunit 8 (EC 1.6.5.3)</td>
</tr>
<tr>
<td>A8IV7J</td>
<td>NUO51</td>
<td>NADH:ubiquinone oxidoreductase 76 kDa subunit</td>
</tr>
<tr>
<td>Q9F6S5</td>
<td>NUO7</td>
<td>NADH:ubiquinone oxidoreductase 49 kDa subunit (EC 1.6.5.3)</td>
</tr>
<tr>
<td>A8HRL8</td>
<td>DLD2*</td>
<td>Dihydrolipoyl dehydrogenase (EC 1.8.1.4)</td>
</tr>
<tr>
<td>A8J766</td>
<td>DLA2*</td>
<td>Dihydrolipoyl dehydrogenase (EC 2.3.1.12)</td>
</tr>
<tr>
<td>A8J776</td>
<td>EAR</td>
<td>Enoyl ACP reductase</td>
</tr>
<tr>
<td>A8J6F4</td>
<td>BC</td>
<td>Biotin carboxylase, acetyl-CoA carboxylase component</td>
</tr>
<tr>
<td>A8J8B0</td>
<td>ALAD</td>
<td>Delta-aminolevulinic acid dehydratase (EC 4.2.1.24)</td>
</tr>
<tr>
<td>A8J7H3</td>
<td>GSA</td>
<td>Glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8)</td>
</tr>
<tr>
<td>A8HNE8</td>
<td>CHIP</td>
<td>Geranylgeranyl reductase (EC 1.3.1.1)</td>
</tr>
<tr>
<td>A8J9J2</td>
<td>POR</td>
<td>Light-dependent protochlorophyllide reductase (EC 1.3.1.33)</td>
</tr>
<tr>
<td>A8J8H3</td>
<td>PSAT</td>
<td>Phosphoserine aminotransferase (EC 2.6.1.52)</td>
</tr>
<tr>
<td>A8J786</td>
<td>HIS7</td>
<td>Imidazole glycerol phosphate synthase (Fragment)</td>
</tr>
<tr>
<td>A8JX6A</td>
<td>AGGPR</td>
<td>N-acetyl-gamma-glutamyl-phosphate reductase</td>
</tr>
<tr>
<td>A8H959</td>
<td>LEU2</td>
<td>Isovaleraldehyde synthase (EC 2.3.3.13) (Fragment)</td>
</tr>
<tr>
<td>A8JX0E</td>
<td>SAHH</td>
<td>S-Adenosylhomocysteine hydrolase (EC 3.3.1.1)</td>
</tr>
<tr>
<td>A8JF4R</td>
<td>OTA1</td>
<td>Ornithine transaminase (EC 2.6.1.13)</td>
</tr>
<tr>
<td>A8J9G0</td>
<td>LEU1</td>
<td>Isovaleraldehyde dehydrogenase, large subunit</td>
</tr>
<tr>
<td>A8J9J7</td>
<td>MCCa</td>
<td>Methylcrotonoyl-CoA carboxylase alpha subunit</td>
</tr>
<tr>
<td>A8J826</td>
<td>HSK1</td>
<td>Homoserine kinase (EC 2.7.1.39)</td>
</tr>
<tr>
<td>A8I73H</td>
<td>ANS1</td>
<td>Anthranilate synthase, alpha subunit (Fragment)</td>
</tr>
<tr>
<td>A8I73J</td>
<td>ASD1</td>
<td>Aspartate semialdehyde dehydrogenase</td>
</tr>
<tr>
<td>A8J726</td>
<td>Chs1</td>
<td>Chorismate synthase (EC 4.2.3.5)</td>
</tr>
<tr>
<td>A8JX0B</td>
<td>AAD1</td>
<td>Acetoxyacetyl decarboxylase</td>
</tr>
<tr>
<td>A8J8S0</td>
<td>OAS7L4</td>
<td>O-acetylsereine sulphohydrolase; cysteine synthase (EC 2.5.1.47)</td>
</tr>
<tr>
<td>A8J3S5</td>
<td>CGS1</td>
<td>Cystathionine gamma-synthase</td>
</tr>
<tr>
<td>A8I3D3</td>
<td>HIS5</td>
<td>Histidin phosphate aminotransferase</td>
</tr>
<tr>
<td>A8I7U7</td>
<td>PGS1</td>
<td>D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)</td>
</tr>
<tr>
<td>D22S47</td>
<td>AIA5</td>
<td>Acetolactate synthase</td>
</tr>
<tr>
<td>A8J8R0</td>
<td>CPS1</td>
<td>Carbamoyl phosphate synthase, large subunit (EC 6.3.5.5)</td>
</tr>
<tr>
<td>P2676S</td>
<td>ARG7</td>
<td>Argininosuccinate lyase (ASAL) (EC 4.3.2.1) (Argininosuccinase)</td>
</tr>
<tr>
<td>A8J305</td>
<td>GLN1</td>
<td>Glutamine synthetase (EC 6.3.1.2)</td>
</tr>
<tr>
<td>A8J8L3</td>
<td>NAGS</td>
<td>N-acetylglutamate synthase (Fragment)</td>
</tr>
<tr>
<td>Category</td>
<td>Gene</td>
<td>Gene</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Other</td>
<td>A8J841</td>
<td>THICb</td>
</tr>
<tr>
<td></td>
<td>A8HM20</td>
<td>CRT2</td>
</tr>
<tr>
<td></td>
<td>Q8LRU1</td>
<td>FER1</td>
</tr>
<tr>
<td>Kinases</td>
<td>A8JD2</td>
<td>ADK3</td>
</tr>
<tr>
<td></td>
<td>A8JH2</td>
<td>-</td>
</tr>
<tr>
<td>Redox proteins</td>
<td>D22472</td>
<td>CAT</td>
</tr>
<tr>
<td></td>
<td>A8HQT2</td>
<td>PD2</td>
</tr>
<tr>
<td></td>
<td>A8I47</td>
<td>GSH1</td>
</tr>
<tr>
<td></td>
<td>A8IV9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A8RSW2</td>
<td>-</td>
</tr>
<tr>
<td>N &amp; S metabolism</td>
<td>A8IB4</td>
<td>GSH1</td>
</tr>
<tr>
<td></td>
<td>A8XF1</td>
<td>XSF1</td>
</tr>
<tr>
<td></td>
<td>A8IA7</td>
<td>MET16</td>
</tr>
<tr>
<td></td>
<td>A8JD1</td>
<td>GDH2</td>
</tr>
<tr>
<td>Cell structure/organisation</td>
<td>A8IBS</td>
<td>APG5</td>
</tr>
<tr>
<td></td>
<td>A8H56</td>
<td>CDC48</td>
</tr>
<tr>
<td></td>
<td>A8IE4</td>
<td>FAP42</td>
</tr>
<tr>
<td></td>
<td>A8FL6</td>
<td>FAP262</td>
</tr>
<tr>
<td></td>
<td>A8IJQ</td>
<td>FAP79</td>
</tr>
<tr>
<td></td>
<td>A8IAV1</td>
<td>iDAS</td>
</tr>
<tr>
<td></td>
<td>P9305</td>
<td>TUBA2</td>
</tr>
<tr>
<td></td>
<td>A8IX20</td>
<td>TUB1</td>
</tr>
<tr>
<td>Stress response</td>
<td>Q9503</td>
<td>HSP70B</td>
</tr>
<tr>
<td></td>
<td>A8IZU</td>
<td>HSP70C</td>
</tr>
<tr>
<td></td>
<td>P5840</td>
<td>HSP70</td>
</tr>
<tr>
<td></td>
<td>A8IT1</td>
<td>HSP90B</td>
</tr>
<tr>
<td></td>
<td>A8J27</td>
<td>CLIPB</td>
</tr>
<tr>
<td></td>
<td>A8HYZ3</td>
<td>HSP708</td>
</tr>
<tr>
<td>DNA &amp; RNA metabolism</td>
<td>A8I3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Q14Z1</td>
<td>PNP</td>
</tr>
<tr>
<td></td>
<td>A8IZS</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A8TI5</td>
<td>ORC1</td>
</tr>
<tr>
<td></td>
<td>A8C71</td>
<td>HBV1</td>
</tr>
<tr>
<td></td>
<td>A8HWY2</td>
<td>HTR14</td>
</tr>
<tr>
<td></td>
<td>A8HYA3</td>
<td>HFO4</td>
</tr>
<tr>
<td></td>
<td>A8JNN9</td>
<td>HFO3</td>
</tr>
<tr>
<td></td>
<td>A8HRZ9</td>
<td>H2A</td>
</tr>
<tr>
<td>Transporters</td>
<td>A8GB0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A8I64</td>
<td>ATPA1</td>
</tr>
<tr>
<td></td>
<td>A8268</td>
<td>HLA3</td>
</tr>
</tbody>
</table>