

ALKALINE MICROALGAE FROM YELLOWSTONE
NATIONAL PARK: PHYSIOLOGICAL AND GENOMIC
CHARACTERIZATION FOR BIOFUEL PRODUCTION

by

Karen Margaret Moll

A dissertation submitted in partial fulfillment

of the requirements for the degree

of

Doctor of Philosophy

in

Microbiology

MONTANA STATE UNIVERSITY

Bozeman, Montana

April 2021

©COPYRIGHT

by

Karen M. Moll

2021

All Rights Reserved

DEDICATION

This dissertation is dedicated to Robert D. Gardner.

ACKNOWLEDGEMENTS

I wish to thank my committee members for their support and guidance resulting in the successful completion of my graduate work: Dr. Brent Peyton, Dr. Matthew Fields, Dr. Joann Mudge, and Dr. Mensur Dlakic and Dr. Robin Gerlach for feedback. I would especially like to take a moment to acknowledge Brent Peyton for his continued support throughout the years. Thank you to Mensur Dlakic for your help with ezTree. I am especially grateful to Donna Negaard for keeping me on track throughout this process. Thank you to all present and past members of the Peyton Lab, especially Rob Gardner, Lisa Kirk, Tisza Bell, Muneeb Rathore, Everett Eustance, and Todd Pedersen. This work could not have been done without help in the lab over the years: Hannah Newhouse, Nathan Murphy, Daniel McDonald, Cansu Bozbiyik, Burcu Orza, and Berrak Erturk. I would also like to thank my brother, Brian, for always being my biggest advocate, my mother and Pamela Cersosimo for your love, guidance, wisdom and support. Thank you to my friends who have been supportive over the years, especially, Karla Sartor and Brian Larson, Anne Rockhold and Sarah Huth. Thank you to everyone at The Center for Biofilm Engineering for your friendships and providing a collaborative work environment. I want to acknowledge everyone at NCGR, especially Joann Mudge, Thiru Ramaraj, Connor Cameron, Callum Bell, Kathy Meyers, Anitha Sundararjan, and Nico Devitt. This work was supported by a NM INBRE Pilot Award, Bridge Funding, Dissertation Completion Awards, TBI, DOE, NSF, and Yellowstone National Park.

TABLE OF CONTENTS

1. INTRODUCTION	1
Background	1
Difference between Prokaryotic and Eukaryotic Genome Projects.....	2
Sequencing Types	3
Sequencing by Synthesis.....	3
454 Pyrosequencing.....	3
Illumina	4
Long-Read Technologies	5
Genome Assembly	5
Scaffolding Approaches.....	5
Combining Technologies	6
What Is a Good Genome Assembly?	7
Eukaryotic Genome Annotation	9
Conclusion	9
Dissertation Overview	10
2. BIODIESEL (MICROALGAE).....	13
Contribution of Authors and Co-Authors	13
Manuscript Information	14
Introduction.....	15
Microalgae	17
Extreme Environments.....	20
Targeting Extremophiles.....	21
Bioprospecting	23
Algae as Biofuels	25
Other Secondary Products.....	31
Take Home Message.....	32
3. CHARACTERIZATION OF NINE NOVEL GREEN ALGAE STRAINS FROM YELLOWSTONE NATIONAL PARK	33
Contribution of Authors and Co-Authors	33
Manuscript Information	34

TABLE OF CONTENTS CONTINUED

Abstract.....	35
Introduction.....	36
Materials & Methods	38
Screening Studies.....	39
In-depth Characterization Studies.....	40
Bicarbonate Addition	40
Determination of Unialgal Strains and Strain Identification	41
Dry Cell Weight.....	41
Nitrate	41
Nile Red Fluorescence	42
Results.....	42
Verification of Unialgal Strains and Strain Identification	43
Doubling Time	46
Biomass Production	49
Highest Lipid Producing Strains.....	53
Discussion.....	56
Growth rate	56
Biomass production	56
Lipid production.....	57
Summary & Conclusions	61
4. DRAFT GENOME FOR A NOVEL, EXTREMOPHILIC, FRESHWATER DIATOM	62
Contribution of Authors and Co-Authors	62
Manuscript Information	63
Abstract.....	64
Introduction.....	65
Methods.....	68
DNA Extraction	68
Whole-genome Sequencing	68
Illumina.....	68
PacBio.....	68

TABLE OF CONTENTS CONTINUED

Assembly Methods.....	69
BUSCO	69
Read Alignments and Validation	70
Structural Annotation.....	70
Functional Annotation	70
K-mer Analysis	71
Concatenated Protein Phylogenetic Tree	71
16S Amplified Sequencing and Analysis	71
RNA Sequencing and Transcriptome Assembly	72
Transcript Alignment and Assembly	72
Results.....	73
Assembly Comparison.....	73
Gene Space Completeness	74
Transcriptome Assembly	77
Assembly Completeness	80
Comparative Genomics.....	81
RGd-1 Genome-based Metabolic Pathway Analysis.....	84
Phycosphere Bacteria.....	93
Discussion	94
Genome Observations.....	94
Metabolic Observations	96
Bacterial Cohabitants	98
Conclusions.....	101
5. GENOME SEQUENCE FOR AN NOVEL BREVUNDIMONAS STRAIN.....	103
Contribution of Authors and Co-Authors	103
Manuscript Information	104
Abstract.....	105
Announcement	105
6. SUMMARY AND FUTURE DIRECTIONS.....	109
Synopsis	109

TABLE OF CONTENTS CONTINUED

Strain Selection	110
Future Directions	111
Future Work	114
Closing	114
References	116
APPENDICES	144
APPENDIX A: CHARACTERIZATION OF NINE NOVEL GREEN ALGAE STRAINS FROM YELLOWSTONE NATIONAL PARK	145
APPENDIX B: RGD-1 GENOME SUPPLEMENTARY DATA	150
High Molecular Weight DNA Extraction	151
BioNano and Assembly Data	152
Transcript Raw Read Data	153
Metabolic Pathways	222
APPENDIX C: DETERMINING THE EFFECTS OF BLUE LIGHT ON THE RGD-1 GROWTH RATE	228
Introduction	229
Background	231
Methods	233
Results and discussion	235
Conclusions	242
APPENDIX D: THE EFFECTS OF ARSENIC SUPPLEMENTATION ON RGD-1 GROWTH RATE AND LIPID ACCUMULATION	244
Introduction	245
Methods	249
Initial Testing	249
Ash Free Dry Weight (AFDW)	250
FAME analysis using GC-MS	251
Determination of the optimal P:As ratio	252
Results	253
Initial testing	254
Arsenate	257

TABLE OF CONTENTS CONTINUED

Discussion.....	261
Summary & Conclusions	263
Sodium arsenite – Supplementary data.....	263
APPENDIX E: STRATEGIES FOR OPTIMIZING BIONANO AND DOVETAIL EXPLORED THROUGH A SECOND REFERENCE QUALITY ASSEMBLY FOR THE LEGUME MODEL, <i>MEDICAGO TRUNCATULA</i>	
Abstract	272
Background	273
Results.....	276
Assembly Continuity	276
Assembly Completeness	278
Gene Space Completeness	279
Joins and Breaks	280
Joins and Breaks in Relation to A17.....	282
Gaps	283
Ordering of Technologies	285
Final Assembly Draft.....	285
Novel sequences revealed by the R108 assembly.....	287
Chromosomal-scale translocation.....	288
Discussion.....	293
Novel Sequence Was Found in the R108 Assembly	293
Technologies Made Similar Continuity Gains and Are Valuable Individually	293
Further Gains Were Made Using Both Technologies.....	294
Join Accuracy Appears to be Higher in Dovetail Compared To BioNano.....	295
Strengths and Weaknesses Dictate Strategy for Ordering Technologies.....	297
Conclusions.....	298
Methods.....	299
PacBio Sequencing and Assembly.....	299
Dovetail.....	300
BioNano	301

TABLE OF CONTENTS CONTINUED

Illumina	302
Transcriptome assembly	303
BUSCO	303
Read alignments.....	304
Structural Annotation.....	304
Identification of structural rearrangements and novel sequences in R108	305
List of Abbreviations:	305
Availability of data and material.....	306
Additional Files.....	306
APPENDIX F: SOURCES AND RE-SOURCES: IMPORTANCE OF NUTRIENTS, RESOURCE ALLOCATION, AND ECOLOGY IN MICROALGAL CULTIVATION FOR LIPID ACCUMULATION	
Abstract.....	309
Introduction.....	310
Nutrient Dependent Lipid Accumulation.....	313
Nitrogen and Phosphorus.....	316
Carbon.....	319
Silicon Limitation	320
Iron Limitation	321
Biofilm Growth.....	322
Ecological Effects	324
Integrating Life-Cycle Analysis.....	326
Conclusion	328
APPENDIX G: DIRECT MEASUREMENT AND CHARACTERIZATION OF ACTIVE PHOTOSYNTHESIS ZONES INSIDE WASTEWATER REMEDIATING AND POTENTIAL BIOFUEL PRODUCING MICROALGAL BIOFILMS	
Abstract.....	332
Introduction.....	333
Materials and methods	335
Laboratory Strains, Culturing Conditions, and Biomass Sampling.....	335
Outdoor Culturing Conditions.	337
Oxygen Microsensor Analysis.....	337

Lipid Analysis	338
Results and discussion	339
Biofilm Cultivation	339
Field-RABR for Wastewater Remediation	341
Nitrogen Depletion in Lab-RABR Samples	347
Biofuel Precursor Production.....	352
Conclusions.....	356
Appendix G: Supplementary Data	356
APPENDIX H: DISSOLVED INORGANIC CARBON ENHANCED GROWTH, NUTRIENT UPTAKE, AND LIPID ACCUMULATION IN WASTEWATER GROWN MICROALGAL BIOFILMS	357
Manuscript Information	358
Abstract	359
Introduction.....	360
Materials and methods	362
Microalgal biofilm culturing and sampling	362
Water quality monitoring.....	363
Oxygen microsensor analysis	364
Biodiesel analysis.....	364
Results and discussion	365
Microalgae growth rate and yield	365
Removal of nitrogen and phosphorus from synthetic wastewater using algal biofilms	366
Microalgal biofilm photosynthesis and coupled respiration	370
Biofuel precursor production	376
Conclusions.....	378
Appendix H: Supplementary Data	379

LIST OF TABLES

Table	Page
Table 2.1 Examples of extremophilic microalgae and their desirable temperature, pH, and salinity conditions with each having at least one environmental condition outside of normal range.	22
Table 3.1 SSU rDNA (18S) DNA concentrations for the extracted, amplified, and purified prior to 454-Pyrosequencing for the nine YNP green algae strains. DNA was quantified using a Qubit fluorometer with a dsDNA BR Assay kit.	43
Table 3.2 Representative sequences for 18S SSU rDNA. Each sequence was BLAST searched for identification. There were three strains that had identical BLAST results with different identifications for 18S (MF1, PGV6, and PGV10-G1), which represents the diverse collection of sequences in NCBI.	44
Table 3.3 BLAST identification of ITS amplicons obtained by Sanger sequencing. The sequence identity was determined by BLAST search for identification. There were two strains that had identical BLAST results with different identifications ITS (PGV6 and WC-1).	45
Table 4.1 Genome assembly statistics for two RGd-1 assembly versions, v. 1.0 and v. 1.5 (with a small percentage of additional long PacBio reads).....	73
Table 4.2 The RGd-1 v.1.0 genome assembly was analyzed using five BUSCO lineages, Eukaryota, Protists, Alveolata/Stramenopiles, Chlorophyta, and Embryophyta. The gene capture percentage was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies.	75
Table 4.3 Gene capture measured by BUSCO. A total of 303 BUSCOs were searched within the eukaryota lineage.	76
Table 4.4 RGd-1 v.1.0 genome assemblies were compared to the other publicly available diatom genome assemblies, <i>P. tricornutum</i> , <i>T. pseudonana</i> , <i>C. cryptica</i> , <i>P. multiseri</i> and, <i>F. cylindrus</i> . The gene capture percent was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies. A total of 303 BUSCOs were searched within the eukaryota lineage.	77
Table 4.5 Genome assembly statistics for two RGd-1 transcriptome assembly versions, de novo and reference-guided using the RGd-1 v.1.0 genome assembly. Each transcriptome assembly shows the statistics pre- and post-filtering for reads ≥ 500 bp.....	79

Table 4.6 Two different RGd-1 transcriptome assemblies were compared; a <i>de novo</i> assembly and reference-guided assembly. The gene capture percent was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies. A total of 303 BUSCOs were searched within the eukaryota lineage.	79
Table 4.7 Whole-genome alignments using BWA mem where the RGd-1 v. 1.0 genome assembly was indexed and the other assembly was queried. The publicly available diatom genomes, <i>C. cryptica</i> , <i>F. cylindrus</i> , <i>P. multiseriis</i> , <i>P. tricornutum</i> , and <i>T. pseudonana</i> were each aligned to RGd-1 on the nucleotide-level.	82
Table 4.8 Identification for 16S amplified sequencing in the RGd-1 culture. Organisms were identified using the 16S RDB Database within CLARK. ^{199, 200} Each organism was calculated for the percentage of all of the categories below (the 9 genera identified and the unknowns) and the percentage classified (the 9 genera excluding the unknowns).	94
Table 5.1 Genome assembly statistics for <i>Brevundimonas</i> sp., strain KM-427. The genome was assembled with Canu as part of an RGd-1 PacBio sequencing project. ¹⁷⁵	106
Table 5.2 Gene capture measured by BUSCO. A total of 148 BUSCOs were searched within the bacteria odb9 lineage.	106
Table A.1 The optimal Nile Red exposure stain times and stain methods for each of the 11 green algae strains. Each strain was exposed to the lipophilic stain, Nile Red, in 20% DMSO and acetone until an optimal stain time was indicated. The stain method that resulted in higher fluorescence was selected as the proper stain method for each strain because that carrier (DMSO or acetone) was able to cross the cell membrane more effectively. ¹	147
Table A.2 The endpoint DCW and doubling times in the air-only and sodium bicarbonate added conditions for each of the 11 strains. Each condition was grown in triplicate.	148
Table A.3 Final DCWs and doubling times for each green algae strain for the control and sodium bicarbonate addition conditions. The DCWs were the average and 95% confidence interval of each triplicate at the time of harvest for each experiment.	149
Table B.1 DNA Extraction (JGI Method). ² The 1.5, 30 and 60 mL headers refer to the container volumes recommended for the DNA extraction volumes. Fifty milliliters were centrifuged and adjusted to ~OD600 1.0 as indicated in step 6. To improve cell wall breakage, mechanical stress was applied with sterile sand, mortar and pestle and, liquid nitrogen. Rather than using isopropanol, the DNA was suspended in molecular grade ethanol in the -20C freezer overnight to improve DNA precipitation.	151

Table B.2 Assembly statistics for BioNano data. RGd-1 biomass was submitted to the Bioinformatics Center at Kansas State for high molecular weight (HMW) DNA extraction and whole genome map assembly. The HMW DNA was digested using the endonuclease, Dnase1 to introduce nicks and create 3' hydroxyl group. DNA polymerase 1 catalyzed the addition of fluorescently labeled Alexa 546 dUTP fluorescent dyes that attached to the nucleotides at the 3' hydroxyl group. 5' to 3' exonuclease activity removed the nucleotides from the 5' phosphoryl terminus of the nick. The labeled and unlabeled nucleotides displaced the excised nucleotides in the original DNA strand. The fluorescently-labeled DNA were visualized using the intercalating dye, YOYO-1. The labeled DNA was added to an IrysChip flow cell, linearized with an electrophoretic current and imaged.....	152
Table B.3 Pfam proteins. Seven Pfam proteins were found in common among the 18 algal genomes that were used for the concatenated protein tree using the ezTree, pipeline.	152
Table C.1 Growth conditions (filter types) and the measured PAR passing through the filter measured with a spectroradiometer (Ocean Optics).	234
Table C.2 Growth conditions (filter types) and doubling times for blue light growth studies. Each condition was grown in duplicate. One replicate in the yellow condition was excluded due to severe clumping.	239
Table D.1 2013 Witch Creek water analyses and B8.7SiS chemical concentrations.	251
Table D.2 Molar Phosphorus and arsenic ratios used in arsenate experiments. ⁶¹	252
Table D.3 Descriptions of arsenate experiments. Seven experiments were performed with different concentrations of sodium arsenate. The first set of experiments was used to determine the optimal As:P ratio for RGd-1. That phosphorus concentration was used for all future experiments with varying As.	253
Table D.4 Doubling times and DCW for the RGd-1 initial testing with sodium arsenate.	257
Table E.1 Number and characteristics of contigs and scaffolds for each of the five assemblies.	277
Table E.2 Characteristics of Input Scaffolds that were Joined by BioNano and/or Dovetail.....	281
Table E.3 Characteristics of the gaps introduced into the assemblies by BioNano and Dovetail. Note, there are no gaps in the Pb only base assembly so it is not included.	283
Table E.4 Assembly Statistics for R108 version 1.0 (PbDtBn PBJelly gap filled) and its input assembly (PbDtBn).	286

Table E.5 R108 v 1.0 assembly characteristics in comparison to the A17 reference assembly. 288

Table F.1 Genera of 56 eukaryotic, photoautotrophs previously studied and reported for the accumulation of lipids. Modified from Breuer et al. (2012).¹⁸⁵ 317

Table G.1 Measurements of areal photosynthesis rates, areal respiration rates and relevant depth scales for the laboratory- and field-RABR cultured biofilms. 340

Table G.2 Mean extractable biofuel precursor weight % and areal concentrations for the laboratory- and field-RABR cultured biofilms (n = 3 with one standard deviation error, or n=2 with range reported as error).....**Error! Bookmark not defined.**

Table G.3 Mean FAME %, weight %, and areal concentration from the laboratory- and field-RABR cultured biofilms. Biomass was directly transesterified to determine total biofuel potential from all fatty acid precursor molecules (extractable and non-extractable) (n=3 with one standard deviation error, or n = 2 with range reported as error)..... 353

Table H.1 Measurements of photosynthetic rates, respiration rates, and relevant depth parameters for laboratory grown microalgal biofilms with and without bicarbonate amendment. 373

Table H.2 Total and percent composition of extractable biofuel precursor weight (%) in laboratory grown algal biofilms with and without bicarbonate amendment. 377

LIST OF FIGURES

Figure	Page
Figure 2.1 YNP diatom strain RGd-1 (left) and YNP green algal WC-1 (right, scale bar 10µm).....	17
Figure 2.2 Each bar represents the fold difference in Nile Red fluorescence intensities at 15 d for each treatment compared to the 2 mM Si control.	18
Figure 2.3 Inputs from thermal hot springs into Witch Creek (research in Yellowstone was conducted under an approved Yellowstone Research Permit [Permit # 5480])......	21
Figure 2.4 Typical heterogeneity of a sampling site containing green algae, diatoms and cyanobacteria (research in Yellowstone was conducted under an approved Yellowstone Research Permit [Permit # 5480])......	23
Figure 2.5 RGd-1 transmitted light (left) and Nile Red fluorescence under epifluorescent light (right).	25
Figure 2.6 Outdoor raceway pond (2000L) at Utah State University, Logan UT.	27
Figure 2.7 Photobioreactor illuminated (Green Wave Energy, Inc.) by artificial light in pilot-scale laboratory setting at Montana State University, Bozeman MT.....	28
Figure 3.1 Outline for algae isolation beginning from field collection to strain characterization. Each strain was streaked for isolation on solid growth medium, grown in liquid growth medium and visualized microscopically at each step to ensure strain isolation.....	39
Figure 3.2 Full-length ITS Sanger results used for strain identification. Each strain (median = 1194 bp), was aligned with Muscle (v3.8.1551) ¹³⁴ and phylogenetic distances were determined using the Maximum Likelihood method with RaxML (version 8.2.12). ¹²⁷ The scale bar represents number of nucleotide changes between strains. The bootstrap values present at the nodes represent the divergence event on a time scale.	46
Figure 3.3 Average doubling times based on the maximal growth rates for each of the eleven strains (replicate 1 and replicate 2). The error bars represent 95% confidence intervals.....	47

LIST OF FIGURES CONTINUED

Figure	Page
Figure 3.4 Growth and lipid accumulation observations for the cultures with the fastest growth rates, WC-1, WC-2b and WC-5 with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C) and Nile Red fluorescence (D) for the fastest doubling times. The error bars represent 95% confidence intervals. For Figures A-C, the series are represented by the following: WC-5 (circles), WC-1 (triangles) and WC-2b (squares). In Figure D1, WC-1 (circles) and WC-2b (triangles) and Figure D2 WC-5 (circles). The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.....	50
Figure 3.5 Growth and lipid accumulation observations for the cultures with the highest biomass production as DCW, PC-3 (circles), and WC-2b (triangles), with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C) and neutral lipids (D). The error bars represent 95% confidence intervals. The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.....	52
Figure 3.6 Average final Nile Red fluorescence for each of the eleven strains (with and without sodium bicarbonate addition). The error bars represent 95% confidence intervals.	53
Figure 3.7 Growth and lipid accumulating observations for the cultures with the highest lipid production, WC-5 (circles) and UTEX 395 (triangles), with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C), and FAMES (D). The error bars represent 95% confidence intervals. The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.....	55
Figure 4.1 Comparison of two draft RGd-1 genome assemblies, v. 1.0 and v. 1.5. The difference between the two assemblies was the inclusion of an additional small PacBio dataset. This figure was generated using MultiQC. ²⁰⁴	74
Figure 4.2 The number trimmed paired-end reads that were uniquely aligned as pairs, had one mate pair uniquely, one mate mapped in multiple locations, the pairs mapped discordantly, the pair-end reads mapped in multiple locations, or neither read aligned. Reads were aligned using HiSat2 and the figure was generated using MultiQC. ²⁰⁴	78
Figure 4.3 A k-mer sweep generated by KAT ³⁴ using a k-mer length of 27. The analysis was performed using the paired-end, 50 bp, Illumina reads.	81

LIST OF FIGURES CONTINUED

Figure	Page
Figure 4.4 The protein sequences from fifteen publicly available genome annotations (1 red alga, 1 brown alga, 11 green algae and 3 diatoms) and RGd-1 were used to construct a concatenated protein using a modified ezTree pipeline. ¹⁹³ RGd-1 was phylogenetically closest to <i>P. tricornutum</i> on a protein-level. The proteins were trimmed using trimAL ¹⁹⁶ and aligned with MAFFT-L-INS-i. ¹⁹⁵ The scale represents the bootstrap values.	83
Figure 4.5 Comparison of <i>P. tricornutum</i> and RGd-1 assemblies based on amino acid sequences. The x-axes contains the scaffolds for the <i>P. tricornutum</i> assembly and the y-axes contains the 520 scaffolds for the RGd-1 genome assembly. Within the mummer packager, promer translates the nucleic acid-based assemblies into amino acids. ²¹¹ Perfectly syntenous assemblies would have a slope of 1.0.....	84
Figure 4.6 Annotated carbon fixation pathway in photosynthetic organisms. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	86
Figure 4.7 Annotated Citrate Acid Cycle pathway. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	87
Figure 4.8 Annotated Glyoxylate and Dicarboxylate Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	87
Figure 4.9 Annotated fatty acid metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	89
Figure 4.10 Annotated Fatty Acid Degradation Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	90
Figure 4.11 Annotated Glycerolipid Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper. ¹⁹¹	92
Figure 4.12 The citric acid cycle with the glyoxylate and dicarboxylate pathway which diverts isocitrate to malate. ²²⁶ The two enzymes, isocitrate lyase and malate synthase modify the citric acid cycle avoiding two decarboxylation steps resulting in the formation of malate from 2 molecules of acetyl-CoA. ²²⁶	97

LIST OF FIGURES CONTINUED

Figure	Page
Figure 4.13 Potential mechanisms of symbiosis between marine diatoms and bacteria. Phytoplankton, such as diatoms may provide dissolved organic carbon (DOC), particulate organic carbon (POC), and other complex algal polysaccharides. The bacteria may supply micronutrients, macronutrients, and vitamins such as B12. ^{160,162-165}	100
Figure 5.1 The <i>Brevundimonas</i> sp. strain, KM-427, genome annotation major features.	107
Figure A.1 Light microscopy images of the nine YNP green algae isolates. (A) PGV-6 (B) PGV8-G1 (C) PGV8-G2 (D) PGV10-G1 (E) PGV10-G2 (F) WC2b (G) WC-5A (H) MF1 and (I) WC-1.	146
Figure B.1 Each sample was analyzed using FastQC and compiled within MultiQC for their unique and duplicate sequence counts. There were a total of nine samples, three culture conditions and three replicates for each condition. The forward (R1) and reverse (R2) reads were analyzed for each sample. Samples A1-A3 had the largest number of unique reads among the sequenced samples and C1-C3 had the least number of unique reads.	153
Figure B.2 This figure indicates the presence of adapter sequence contamination in sample A1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	154
Figure B.3 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A1-R1. Here, 23.67% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	154
Figure B.4 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	155
Figure B.5 Quality scores across the positions of the 150 bp reads for sample A1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	155

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.6 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5'end of the reads. ³	156
Figure B.7 The per sequence GC content for sample A1-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	156
Figure B.8 The per sequence quality score for sample A1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	157
Figure B.9 The distribution of the sequence lengths for sample A1-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.	157
Figure B.10 This figure indicates the presence of adapter sequence contamination in sample A1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	158
Figure B.11 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A1-R2. Here, 29.88% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	158
Figure B.12 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	159

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.13 Quality scores across the positions of the 150 bp reads for sample A1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	159
Figure B.14 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	160
Figure B.15 The per sequence GC content. The x- and y-axes represent the %GC content per read and read counts, respectively for sample A1-R2. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	160
Figure B.16 The per sequence quality score for sample A1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	161
Figure B.17 The distribution of the sequence lengths for sample A1-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	161
Figure B.18 This figure indicates the presence of adapter sequence contamination in sample A2-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	162
Figure B.19 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A2-R1. Here, 26.72% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	162

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.20 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A2-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	163
Figure B.21 Quality scores across the positions of the 150 bp reads for sample A2-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	163
Figure B.22 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A2-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	164
Figure B.23 The per sequence GC content for sample A2-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	164
Figure B.24 The per sequence quality score for sample A2-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	165
Figure B.25 The distribution of the sequence lengths for sample A2-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.	165
Figure B.26 This figure indicates the presence of adapter sequence contamination in sample A2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	166

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.27 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A2-R2. Here, 33.26% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	166
Figure B.28 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	167
Figure B.29 Quality scores across the positions of the 150 bp reads for sample A2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	167
Figure B.30 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	168
Figure B.31 The per sequence GC content for sample A2-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	168
Figure B.32 The per sequence quality score for sample A2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	169
Figure B.33 The distribution of the sequence lengths for sample A2-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	169

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.34 This figure indicates the presence of adapter sequence contamination in sample A3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	170
Figure B.35 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A3-R1. Here, 19.32% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	170
Figure B.36 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	171
Figure B.37 Quality scores across the positions of the 150 bp reads for sample A3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	171
Figure B.38 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5'end of the reads. ³	172
Figure B.39 The per sequence GC content for sample A3-R1. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	172
Figure B.40 The per sequence quality score for sample A3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	173

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.41 The distribution of the sequence lengths for sample A3-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	173
Figure B.42 This figure indicates the presence of adapter sequence contamination in sample A3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	174
Figure B.43 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A3-R2. Here, 23.64% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	174
Figure B.44 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	175
Figure B.45 Quality scores across the positions of the 150 bp reads for sample A3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	175
Figure B.46 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	176
Figure B.47 The per sequence GC content for sample A3-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.	176

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.48 The per sequence quality score for sample A3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	177
Figure B.49 The distribution of the sequence lengths for sample A3-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	177
Figure B.50 This figure indicates the presence of adapter sequence contamination in sample B1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	178
Figure B.51 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B1-R1. Here, 29.46% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	178
Figure B.52 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	179
Figure B.53 Quality scores across the positions of the 150 bp reads for sample B1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	179
Figure B.54 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5'end of the reads. ³	180

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.55 The per sequence GC content for sample B1-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	180
Figure B.56 The per sequence quality score for sample B1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	181
Figure B.57 The distribution of the sequence lengths for sample B1-R1. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	181
Figure B.58 This figure indicates the presence of adapter sequence contamination in sample B1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	182
Figure B.59 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B1-R2. Here, 24.99% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	182
Figure B.60 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	183
Figure B.61 Quality scores across the positions of the 150 bp reads for sample B1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	183

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.62 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5'end of the reads. ³	184
Figure B.63 The per sequence GC content for sample B1-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	184
Figure B.64 The per sequence quality score for sample B1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	185
Figure B.65 The distribution of the sequence lengths for sample B1-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	185
Figure B.66 This figure indicates the presence of adapter sequence contamination in sample B2-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	186
Figure B.67 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B2-R1. Here, 14.12% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	186
Figure B.68 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B2-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	187

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.69 Quality scores across the positions of the 150 bp reads. The x- and y-axes represent the quality scores and position within the read for sample B2-R1. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	187
Figure B.70 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B2-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	188
Figure B.71 The per sequence GC content for sample B2-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	188
Figure B.72 The per sequence quality score for sample B2-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	189
Figure B.73 The distribution of the sequence lengths for sample B2-R1. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	189
Figure B.74 This figure indicates the presence of adapter sequence contamination in sample B2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	190
Figure B.75 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B2-R2. Here, 14.12% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	190

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.76 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	191
Figure B.77 Quality scores across the positions of the 150 bp reads for sample B2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	191
Figure B.78 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	192
Figure B.79 The per sequence GC content for sample B2-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	192
Figure B.80 The per sequence quality score for sample B2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	193
Figure B.81 The distribution of the sequence lengths for sample B2-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	193
Figure B.82 This figure indicates the presence of adapter sequence contamination in sample B3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	194

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.83 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B3-R1. Here, 20.06% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	194
Figure B.84 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	195
Figure B.85 Quality scores across the positions of the 150 bp reads for sample B3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	195
Figure B.86 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	196
Figure B.87 The per sequence GC content for sample B3-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	196
Figure B.88 The per sequence quality score for sample B3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	197
Figure B.89 The distribution of the sequence lengths for sample B3-R1. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	197

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.90 This figure indicates the presence of adapter sequence contamination in sample B3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	198
Figure B.91 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B3-R2. Here, 15.98% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	198
Figure B.92 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.....	199
Figure B.93 Quality scores across the positions of the 150 bp reads for sample B3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	199
Figure B.94 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5'end of the reads. ³	200
Figure B.95 The per sequence GC content for sample B3-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.	200
Figure B.96 The per sequence quality score for sample B3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	201

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.97 The distribution of the sequence lengths for sample B3-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	201
Figure B.98 This figure indicates the presence of adapter sequence contamination in sample C1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	202
Figure B.99 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C1-R1. Here, 4.65% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	202
Figure B.100 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.	203
Figure B.101 Quality scores across the positions of the 150 bp reads for sample C1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	203
Figure B.102 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	204
Figure B.103 The per sequence GC content for sample C1-R1. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination.	204

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.104 The per sequence quality score for sample C1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	205
Figure B.105 The distribution of the sequence lengths for sample C1-R1. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	205
Figure B.106 This figure indicates the presence of adapter sequence contamination in sample C1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	206
Figure B.107 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C1-R2. Here, 8.8% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	206
Figure B.108 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.	207
Figure B.109 Quality scores across the positions of the 150 bp reads for sample C1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	207
Figure B.110 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	208

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.111 The per sequence GC content for sample C1-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean % GC = 47 indicates bacterial contamination.	208
Figure B.112 The per sequence quality score for sample C1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	209
Figure B.113 The distribution of the sequence lengths for sample C1-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	209
Figure B.114 This figure indicates the presence of adapter sequence contamination in sample C2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	210
Figure B.115 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C2-R2. Here, 7.08% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	210
Figure B.116 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.	211
Figure B.117 Quality scores across the positions of the 150 bp reads for sample C2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	211

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.118 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	212
Figure B.119 The per sequence GC content for sample C2-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination. The sharp peak may indicate overrepresented bacterial reads.	212
Figure B.120 The per sequence quality score for sample C2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	213
Figure B.121 The distribution of the sequence lengths for sample C2-R2. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	213
Figure B.122 This figure indicates the presence of adapter sequence contamination in sample C3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	214
Figure B.123 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C3-R1. Here, 29.88% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	214
Figure B.124 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.	215

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.125 Quality scores across the positions of the 150 bp reads for sample C3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	215
Figure B.126 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	216
Figure B.127 The per sequence GC content for sample C3-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination.	216
Figure B.128 The per sequence quality score for sample C3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	217
Figure B.129 The distribution of the sequence lengths for sample C3-R1. The x- and y- axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	217
Figure B.130 This figure indicates the presence of adapter sequence contamination in sample C3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.	218
Figure B.131 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C3-R2. Here, 8.93% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.	218

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.132 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.	219
Figure B.133 Quality scores across the positions of the 150 bp reads for sample C3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.....	219
Figure B.134 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads. ³	220
Figure B.135 The per sequence GC content for sample C3-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination. The sharp peak may indicate overrepresented bacterial reads.	220
Figure B.136 The per sequence quality score for sample C3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.	221
Figure B.137 The distribution of the sequence lengths for sample C3-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.....	221
Figure B.138 The glycolysis/gluconeogenesis metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	222
Figure B.139 The pyruvate metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	223

LIST OF FIGURES CONTINUED

Figure	Page
Figure B.140 The fatty acid degradation metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	224
Figure B.141 The glycerolipid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	225
Figure B.142 The α -linoleic acid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	226
Figure B.143 The arachidonic acid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper. ⁴	227
Figure C.1 RGd-1 cellular morphologies as imaged using field emission – scanning electron microscopy. The RGd-1 morphology in 2009 (left), and different cell morphologies in 2014 (middle and right).	230
Figure C.2 The light intensity at Witch Creek (late morning August 2012, left) and laboratory fluorescent grow lights (right). Two measurements were taken in the field (1807 and 1828 $\text{uW cm}^{-2} \text{nm}^{-1}$) and three measurements were taken for the MSU lab light systems (421, 412 and 379 $\text{uW cm}^{-2} \text{nm}^{-1}$).	235
Figure C.3 The control (without a color filter) spectroradiometer measurement at 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.	237
Figure C.4 Spectroradiometer measurement (Ocean Optics) for the Rosco filter #313 (light yellow) with low blue intensity using 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.	237
Figure C.5 Spectroradiometer measurements (Ocean Optics) for the Rosco filter #6 (yellow) high blue intensity using 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.	238
Figure C.6 Spectroradiometer measurements (Ocean Optics) for the Rosco filter #384 (blue) very high blue intensity, and low intensity for other wavelengths and 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.	238

LIST OF FIGURES CONTINUED

Figure	Page
Figure C.7 Cell concentrations for the 4 blue light conditions tested, the no light filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.	239
Figure C.8 Total chlorophyll concentrations (mg/mL) for each of the 4 blue light conditions tested, the no-filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.	241
Figure C.9 The Nile Red fluorescence (rfu) for the 4 blue light conditions tested, the no-filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.	241
Figure D.1 Speciation of As(III) and As(IV) across pH ranges -2 to 14 in water. ^{57, 58}	249
Figure D.2 The Nile Red fluorescence for the initial testing with sodium arsenate. The error bars represent the standard deviation of the mean.	255
Figure D.3 The doubling times for the initial testing with sodium arsenate. Two conditions did not grow; Witch Creek Water with Bold's additions at pH 8 + EDTA. This condition was repeated to verify the result. The error bars represent the standard deviation of the mean.	256
Figure D.4 Cell counts for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.	258
Figure D.5 The pH for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.	259
Figure D.6 The total Nile Red fluorescence for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.	259
Figure D.7 The doubling times for the different arsenate concentrations tested. The error bars represent the variance resulting from an ANOVA (2-factor without replacement) analysis. The error bars represent the standard deviation of the mean.	260
Figure D.8 The average ash-free dry weights for the different arsenate concentrations tested. The error bars represent the variance resulting from an ANOVA (2-factor without replacement) analysis. The error bars represent the standard deviation of the mean.	261

LIST OF FIGURES CONTINUED

Figure	Page
Figure D.9 The doubling time for RGd-1 cultures grown in the presence of sodium arsenite in place of sodium arsenate at varying concentrations.	264
Figure D.10 The Nile Red fluorescence for RGd-1 grown in the presence of sodium arsenite instead of sodium arsenite at varying concentrations.	264
Figure E.1 Synteny alignment of partial chromosomes 4 and 8 between A17 and R108 confirms rearrangement of the long arms of the chromosomes.	289
Figure E.2 Synteny alignment of partial A17 chromosomes 4 and 8 against syntenic regions in the R108 Illumina-based assembly (top panel), PacBio-based assembly (Pb, middle panel) as well as the gap-filled PbDtBn (v1.0) assembly (bottom panel).	291
Figure E.3 Schematic of the rearrangement between chromosomes 4 and 8 in A17 (left) compared to R108 (right). Green segments indicate homology to A17's chromosome 4 while blue segments indicate homology to A17 chromosome 8. Red segments indicate sequences not present in the A17 reference). Breakpoint 1 (br1) is pinpointed to a 104 bp region (chr4:39,021,788-39,021,891) and includes a 100 bp gap. Breakpoint 2 (br2) is pinpointed to a 7,665 bp region (chr8:33,996,308-34,003,972) and includes a 7,663 bp gap. Breakpoint 3 (br3) is pinpointed to a 708 bp region (chr8: 34,107,285-34,107,992) and includes a 100 bp gap. Breakpoint 4 is pinpointed to a 277 bp region (chr8:34,275,249-34,275,525) and includes a 100 bp gap).	292
Figure F.1 The biological recycling of carbon, nitrogen, and phosphorus to harvest fuel and food linked to sunlight to reduce net consumption of N and P and net production of C.	313
Figure F.2 Hypothetical performance curve for an increasingly perturbed (i.e., stressed) microalgal system being used to produce photoautotrophic biomass and/or lipids. Adapted from Odum et al. (1979). ¹⁷⁵	314
Figure F.3 Primary stages and (alternative processes) in the microalgae to fuel production process.	326
Figure G.1 Representative photographs for: (A) the field-RABR and (B) lab-RABR culturing systems designed for algal biofilm culturing (insert shows cross-sectioned excised cotton cord substratum with biofilm growth). Note the 'top' and 'bottom' biofilm orientation corresponding to the outer and inner sections of the field-RABR wheel, respectively.	334

LIST OF FIGURES CONTINUED

Figure	Page
Figure G.2 Field-RABR: dissolved oxygen microprofiles measured in the light extending from the surface for biofilms grown on the (A) outer wheel surface and (B) inner wheel surface; dissolved oxygen microprofiles measured in the dark for biofilms grown on the (C) outer wheel surface and (D) inner wheel surface; and photosynthesis profiles extending from the surface for biofilms grown on the (E) outer wheel surface and (F) inner wheel surface. Note that the biofilm surface position (depth = 0) is approximated by the position at which oxygen responses were measureable (subject to $\pm 25 \mu\text{m}$ error or $\pm 100 \mu\text{m}$ error for the photosynthesis profiles where each data point is a representative gross volumetric photosynthesis rate from 2-3 replicates.) and individual data points represent the mean values from 3-4 replicate profiles in both light and dark conditions. Error bars represent plus or minus one standard deviation. Dotted lines indicate the photic-zone termination depth, estimated from the light:dark shift method. Note the scale change on the x-axis.....	342
Figure G.3 lab-RABR: dissolved oxygen microprofiles measured in the light extending from the surface for biofilms grown in (A) nitrate replete and (B) nitrate deplete conditions; dissolved oxygen microprofiles measured in the dark for biofilms grown in (C) nitrate replete and (D) nitrate deplete conditions; and photosynthesis profiles extending from the surface for biofilms grown in (E) nitrate deplete and (F) nitrate deplete. Note that the biofilm surface position (depth = 0) is approximated by the position at which oxygen responses were measurable (subject to $\pm 25 \mu\text{m}$ error or $\pm 100 \mu\text{m}$ error for the photosynthesis profiles where each data point is a representative gross volumetric photosynthesis rate from 2-3 replicates.) and individual data points represent the mean values from 3-4 replicate profiles in both light and dark conditions. Error bars represent plus or minus one standard deviation. Dotted lines indicate the photic-zone termination depth, estimated from the light:dark shift method. Note the scale change on the x-axis.	349
Figure H.1 Growth curve from log transformed data showing the exponential phase (day 3-10) and stationary phase (day 11 -18). Insert: Equations and R ² describing the exponential phase for both biofilms with and without bicarbonate amendment.	365
Figure H.2 Growth curves for attached and suspended microalgae (A) and dissolved inorganic carbon (DIC) concentrations (B) in laboratory-RABRs amended with bicarbonate and without bicarbonate addition. Error bars for algal biofilm yield and DIC measurements represent standard deviation (n=4). Error bars for suspended growth represent range (n=2). Vertical dotted lines represent end of 5 day hydraulic retention time.	367
Figure H.3 Ammonium, nitrate, nitrite, and phosphate ion concentrations in medium amended with bicarbonate and without bicarbonate addition. Error bars represent range for (n=2). Vertical dotted lines represent end of 5 day hydraulic retention time.....	369

LIST OF FIGURES CONTINUED

Figure	Page
Figure H.4 Steady state oxygen microprofiles for illuminated algal biofilms under nitrogen replete (A) and nitrogen deprived (B) conditions. Error bars represent standard deviation of replicate profiles (n=3); steady state oxygen microprofiles in the dark for algal biofilms under nitrogen replete (C) and nitrogen deprived (D) conditions. Error bars represent standard deviation of replicate profiles (n=3); and representative photosynthesis profiles for algal biofilms under nitrogen replete (E) and nitrogen deprived (F) conditions. Zero depth (surface) is at the algal biofilm/air interface.	371
Figure H.5 Total FAMES and free fatty acid composition of the FAMES. A: Mean percent FAME (w/w), B: percent lipids (w/w), C: areal concentration (g m ⁻²). Error bars represent range (n=2). ND and NR represent nitrogen deprived and replete algal biofilms, respectively.	378

GLOSSARY

YNP	Yellowstone National Park
FAME	Fatty acid methyl ester
TAG	Triacylglycerol
DIC	Dissolved inorganic carbon
DCW	Dry cell weight
GC-MS	Gas chromatography mass spectrometry
PAR	Photosynthetically active radiation
BP	Biofuel potential
rfu	Relative fluorescence units
AFDW	Ash free dry weight
NGS	Next generation sequencing
ASP	Aquatic Species Program
GB	Gigabyte
UTR	Untranslated region
EST	Expressed sequence tag
ZMW	Zero mode wave
SNA	Single nucleotide addition
PPi	Pyrophosphate
SBS	Sequencing by synthesis
TIRF	Total internal reflection fluorescence
HMW	High molecular weight (DNA)
OUT	Operational taxonomic units

ABSTRACT

Alternatives are needed to avoid future economic and environmental impacts from continued exploration, harvesting transport, and combustion of conventional hydrocarbons resulting in a rise of atmospheric CO₂. Microalgae, including diatoms, are eukaryotic photoautotrophs that can utilize inorganic carbon (*e.g.*, CO₂) as a carbon source and sunlight as an energy source, and many microalgae can store carbon and energy in the form of neutral lipids. In addition to accumulating useful precursors for biofuels and chemical feed-stocks, the use of autotrophic microorganisms can further contribute to reduced CO₂ emissions through utilization of atmospheric CO₂. Most microalgal biofuel research has focused on green algae. However, there are good reasons to consider diatoms for biofuel research. Diatoms are responsible for approximately 40% of marine primary productivity, are important in freshwater systems, and are known to assimilate 20% of global CO₂. Identification and implementation of factors that can contribute to rapid growth will minimize inputs and production costs, thus improving algal biofuel viability. Nine green algae strains that were isolated from Witch Creek, Yellowstone National Park, were compared to two culture collection strains (PC-3 and UTEX395) for growth rates, dry cell weights and lipid accumulation. The strains exhibiting the fastest growth rates were WC-5, WC-1 and WC-2b. The culture collection strain was the best biomass producer and WC-5 and UTEX395 were the most productive for lipid. Based on the growth rates and lipid content, the best strains for biodiesel production were WC-1 and WC-5. In addition to the green algae strains, diatom strain, RGd-1 has previously been found to accumulate 30-40% (w/w) triacylglycerol and 70-80% (w/w) fatty acid methyl esters that can be transesterified into biodiesel. The RGd-1 was sequenced via Illumina 2x50 and PacBio RSII reads and genome comparisons revealed that the RGd-1 genome is significantly divergent from other publicly available genome sequences. RGd-1 was found to have nearly complete metabolic pathways for fatty acid elongation using acetyl-CoA in the mitochondrion or malonyl-CoA in the cytoplasm. The ability to switch between two different starting substrates may confer an advantage for fatty acid and neutral lipid biosynthesis. Further, RGd-1 was found to use the glyoxylate shunt as part of its central carbon metabolism. This carbon conservation pathway may potentially explain why RGd-1 is able to produce high concentrations of lipids. Using Illumina[®] MiSeq sequencing it was possible to obtain thorough community analysis of bacteria associated with RGd-1 in culture. Nine primary taxa were identified and further research will elucidate their roles as potential phycosphere bacteria that may have specific functional roles that contribute to RGd-1 health. With long-range PacBio reads, RGd-1 was found to have a potential bacterial symbiont, *Brevundimonas* sp.

1

CHAPTER ONE

INTRODUCTION

Background

As the world population continues to increase, especially in countries with high energy needs such as the United States, China, and India, transportation fuel demand will increase,¹ likely exceeding production.² However, the search for alternative fuels has been impeded by reluctance to invest in technologies that cannot compete with low oil prices. For instance, the U.S. Department of Energy's "Aquatic Species Program" (ASP), after screening more than 3,000 green algae and diatom species as microalgal biofuel candidates, was cut because the cost of algal biofuel production was higher than that of petrofuels production.³

Improving algal biofuel viability is beneficial for renewable fuel resources and minimizing US dependence on non-renewable fossil fuels. Regardless of current market conditions and availability of conventional petroleum sources, alternatives are needed to avoid future economic and environmental impacts from continued exploration and harvesting of conventional hydrocarbons. Microalgae, including diatoms, are eukaryotic photoautotrophs that can utilize inorganic carbon (e.g., CO₂) as a carbon source and sunlight as an energy source, and many microalgae can store carbon and energy in the form of neutral lipids. In addition to accumulating useful precursors for biofuels and chemical feed-stocks, the use of autotrophic microorganisms can further contribute to reduced CO₂ emissions through utilization of atmospheric CO₂.⁴ For these reasons, microalgae have been studied in the context of lipid accumulation for over 50 years. Even with conservative estimates of lipid accumulations (25–30% [w/w]), microalgae could replace 50% of U.S. transportation fuel need with an area

equivalent to 3% of U.S. arable cropland.^{5, 6} With revived interest in microalgal fuels,⁷⁻⁹ significant fundamental and applied research is needed to fully maximize biomass and biochemical production for biofuels and renewable biochemicals.⁴

Most microalgal biofuel research has focused on green algae, however, diatoms are unique photoautotrophs that also can produce lipids. Diatoms are responsible for approximately 40% of marine primary productivity and are known to assimilate 25–45% of global CO₂.¹⁰⁻¹² To provide perspective on how much CO₂ diatoms fix per year, the Amazon Rainforest fixes approximately 2 gigatons of CO₂ from the atmosphere each year compared to the 50 gigatons fixed each year by diatoms globally.^{13, 14} Further, in addition to high lipid accumulation, diatoms can accumulate high concentrations of other carbonaceous compounds useful for production of renewable fuels and high value coproducts.¹⁵

Identification and implementation of factors that will contribute to rapid algal/diatom growth while minimizing inputs into algal growth systems and production costs will improve algal biofuel viability. Here, I focused on screening 11 green algae strains for biodiesel production while sequencing and assembling the RGd-1 genome to identify potential novel genes and pathways that are responsible for RGd-1 accumulating high concentrations of lipids. However, given some of the unique characteristics of diatoms and the low number of available diatom genomes, the sequencing, assembly, and analysis of diatom genomes is not without challenges.

Difference between Prokaryotic and Eukaryotic Genome Projects

Eukaryotic genomes are inherently more complex than bacterial and archaeal genomes. The smaller genomes of bacteria and archaea are relatively simple to sequence and assemble. Unlike bacteria and archaea, eukaryotes contain introns, long repetitive regions that are difficult

to sequence and assemble and, have alternative gene-splicing.¹⁶ The final mRNA will contain a selection of exons that were differentially expressed under different growth conditions and alternative splicing represents different isoforms of a gene.¹⁶ Furthermore, eukaryotic genomes may have multiple chromosomes and be polyploid, which can make the creation of one reference genome assembly difficult.¹⁶ As a result, there are extra challenges associated with creating a reference assembly for eukaryotic genomes, especially if there is not a reference available for a similar strain.

Sequencing Types

Sequencing by Synthesis

Next generation sequencing was revolutionized by the high-throughput, sequencing by synthesis (SBS) technologies. In SBS, the addition of each labelled nucleotide is tracked. Here we will focus on two types of SBS technologies; pyrosequencing (454 Pyrosequencing) and sequencing by reversible termination (Illumina).¹⁷

454 Pyrosequencing

454 Pyrosequencing was the first Next-Generation Sequencing type.¹⁸ As a bioluminescent method, light is produced following the release of an inorganic phosphate using a single-nucleotide addition (SNA) method, where pyrophosphate (Ppi) is released and is converted to ATP by ATP sulphurylase using adenosine 5'phosphosulfate.¹⁹ In the presence of ATP, luciferase converts luciferin to oxyluciferin to generate the light.¹⁹ Once dNTPs have been incorporated, DNA polymerase extends the primer and pauses.¹⁷ DNA synthesis is continued when the next round of dNTPs is dispensed. The DNA sequence is determined by the order of the bases and intensity of the light produced by the addition of each base.¹⁷ In 454 Pyrosequencing,

there is only one signal that indicates the addition of a nucleotide. Therefore, base must be added individually at each position. The light is imaged with a charge-coupled device (ccd).¹⁷

Insertions and deletions are the most common error type and homopolymersequencing where the addition of five to six identical bases cannot be detected accurately.^{17, 19} Overall, the accuracy of 454 Pyrosequencing is at least 99%.²⁰

Illumina

Illumina is perhaps the most successful of the short read technologies. Illumina uses cyclic reversible termination (CRT) technology for nucleotide incorporation, fluorescence imaging, and cleavage.^{17, 18} DNA fragments are clonally amplified by bridge PCR, or solid-phase amplification, where forward and reverse primers are attached to the flow cell.^{17, 19} First, DNA polymerase incorporates a fluorescently-labeled nucleotide that is complementary to the template DNA that becomes temporarily suspended from incorporating additional bases due to the blocked by 3'-O-azidomethyl group ribose-3'OH groups. Once the base has been incorporated, the remaining unincorporated fluorescently-labeled nucleotides are washed away. Each of the four bases is labeled consistently with a specific color. The incorporated nucleotide is imaged to determine which base was incorporated, after which, the terminating and fluorescently-labeled groups are removed. Each of the four colors is detected by total internal reflection fluorescence (TIRF) that utilizes two to four lasers, depending on the sequencer type^{17, 18}. Regeneration of the 3'OH group occurs using a reducing agent such as tris(2-carboxyethyl)phosphine (TCEP)¹⁷. Substitutions are the most common error with this chemistry.

Long-Read Technologies

Pacific Biosciences. Unlike the chain termination method utilized by Illumina, Pacific BioSciences (PacBio) utilizes real-time, continuous base-addition measurements. In PacBio chemistry, DNA polymerase is attached to the bottom of zero-mode wave (ZMW) detectors to continuously measure the addition of each base in real time. Following the addition of a nucleotide, the phosphate group containing a fluorophore is cleaved, resulting in fluorescence.¹⁷ Once the hexaphosphate group is cleaved off, the fluorescent signal is quenched, ultimately to background levels. Similar to Illumina technology, each fluorophore corresponds to a different nucleotide. This NGS platform is known to incur significant sequencing errors (83%).¹⁷ However, with sequencing depth and error correction, PacBio reads have been found to be 99% correct. The PacBio HiFi chemistry to the PacBio Sequel is highly accurate at 99% and requires less coverage, approximately 20X coverage compared to 100X coverage needed on the RSII system.²¹⁻²⁴

Genome Assembly

Scaffolding Approaches

In recent years, there has been a renewed interest in optical-mapping technology to improve scaffolding for genome assemblies. BioNano Genomics offers long-read, low-resolution scaffolds generated by one or more restriction endonucleases,¹⁸ a method that is independent of sequencing. This technology employs restriction site mapping, where restriction enzymes introduce fluorophores into restriction sites. The restriction enzymes introduce nicks at specific motifs into double-stranded DNA, creating a 3'OH. DNA polymerase 1 catalyzes the addition of fluorescently labeled Alexa 546 dUTP fluorescent dyes that conjugate to the 3'OH that was

introduced to the double-stranded DNA by the restriction enzyme. The 5'-3' exonuclease activity removes nucleotides from the 5' phosphoryl terminus of the nick. Labeled and unlabeled nucleotides replace excised nucleotides in the sequences, displacing the original DNA strand. To image the long DNA strands, DNA is labeled using the intercalating dye, YOYO-1. The high-molecular-weight DNA fragments are aligned based on the position of the fluorescent restriction sites, that appear as “dots on a string”, to ultimately create a “consensus genome map”.^{25, 26} A reference assembly is then digested in-silico using the same restriction enzyme that was used to create the consensus genome map. The long-reads can improve scaffolding or indicate any erroneously assembled scaffolds such as inversions, insertions, deletions, and gaps regions in comparison to the reference assembly.^{25, 26}

The labeled DNA is linearized on an IrysChip flow cell once electrophoretic movement unravels the DNA. Once the DNA is linear, the current is shut off and the DNA is imaged. Read sizes may vary from hundreds of kilobases to megabase-sized reads.^{25, 26} Once imaging has finished, molecules are flushed and the process is repeated, producing several gigabyte (GB) per hour. BioNano technology is a high throughput technology that can process thousands of parallel channels. Within each channel, only one long strand may be linearized at a time. Once linearized, the Irys CCD detector captures the images.

Combining Technologies

A recent assembly for the domestic goat (*Capra hircus*) used a combined strategy with PacBio, Illumina short-reads, BioNano maps, and Phase Genomics Hi-C chromatin interaction maps.²⁷ Initial contigs were assembled using the PacBio reads, which were further scaffolded and error-corrected by the BioNano maps and finally combined with the Hi-C chromatin maps. Gaps were filled using PacBio reads, and the final assembly was polished using the Illumina

reads resulting in 2.92 Gb assembly with a contig N50 = 19Mb and scaffold N50 = 87 Mb. This combination resulted in the best assembly.

Moll et al. (2016) used a similar approach to generate an assembly for the *Medicago truncatula*, a model organism (clover) that is used to study alfalfa.²⁸ The methods were similar to the goat assembly; however, instead of using Illumina reads for error correction or as part of the genome assembly, Dovetail Genomics was used, a type of Hi-C mapping that uses Illumina reads with the Chicago library preparation. Here, long-range information was generated with high accuracy using the Dovetail Chicago Library preparation method that uses restriction endonucleases on high-molecular-weight DNA.^{28, 29}

What Is a Good Genome Assembly?

There are a number read types (e.g. Illumina or PacBio), read length, quality and library preparation that may contribute to an assembly.¹⁷ While there typically is not a standardized protocol for the steps required for a genome assembly, the following measures indicate the assembly quality:

(1) Short read alignment. Through validation by the alignment of short-reads against the genome assembly.³⁰ Generally, the higher the alignment percentage ($\geq 80\%$), the greater that assembly can capture the high-fidelity reads. A lower alignment may indicate a fragmented assembly that may cause problems with gene calling, as evidenced by BUSCO (see below).

(2) N50. The minimum contig length that represents the assembly 50% of the genome.

While there is no set standard for the N50 required for a good genome, the higher the N50, the more contiguous the genome assembly.¹⁶ With long-reads generated from

PacBio and scaffolding technologies such as BioNano,²⁵ it is possible to get N50s in the megabase range.

- (3) Gene capture. Programs such as BUSCO (Benchmarking Universal Single-Copy Orthologs) assume that orthologs are present for at least 90% of the organisms within a lineage.³¹ When a genome assembly is tested against a lineage, it is possible to determine the number of single-copy, duplicated, missing, or fragmented orthologs in an assembly.³¹ Therefore, it is of utmost importance to choose a lineage that is relevant to the organism represented in the assembly.
- (4) Kmer analysis. Kmers are subsequences, length K that can provide insight into a genome assembly such as genome complexity, ploidy, heterozygosity, repeats, and contamination.³²⁻³⁴ A kmer profile measures how often kmers of each length occur within the short, unassembled reads.³³ A comparison of kmer analyses between the reads and the genome assembly can be used to determine what fraction of reads were incorporated into the assembly.
- (5) Percent GC. The GC content can be an indicator of sample purity or contamination. Two or more GC% peaks may be an indicator of more than one organism represented in the reads, e.g. 43% and 58%. Alternatively, the extra peaks may be over-represented organellar reads.

Using the criteria outlined above, it is possible to determine whether different experimental strategies are required for strain isolation, additional sequencing is needed to improve the assembly or a different sequencing approach is needed to account for a more complex community than anticipated.

Eukaryotic Genome Annotation

Eukaryotic genome annotation using sequence assembly alone yields less accurate results. The reasons are multifactorial: the intronic nature of eukaryotic genomes, alternative splicing, and LTR retrotransposons that may resemble open-reading frames. Unlike prokaryotes, eukaryotic *ab initio* gene prediction using sequence assembly alone cannot identify intron/exon boundaries, untranslated regions (UTR), or alternative splice sites. Further, regions of long-terminal-repeat (LTR) retrotransposons can be identified as protein-coding sequences.¹⁶ Evidence-based methods, such as expressed sequence tags (EST), RNA-seq, or proteomics, are required to identify these sites. For example, approximately 86% of the *P. tricornutum* genome assembly annotation was supported by the 130,000 ESTs acquired from 16 different growth conditions.³⁵

Using whole-genome alignments, there is significant divergence between the diatoms whose genomes are available. For instance, the *P. tricornutum* and *T. pseudonana* (both marine diatoms) genomes are 57% identical on the nucleotide level.³⁶ While it is expected that there will be greater sequence similarity between RGd-1 and *P. tricornutum* due to their pennate cell morphology, we do not expect the genome wide-sequence similarity to be high enough to use *P. tricornutum* or *T. pseudonana* transcriptome or protein files to facilitate genome annotation.

Conclusion

At present, there are only two published, publicly available diatom genomes (*Phaeodactylum tricornutum* and *Thalassiosira pseudonana*) with quality assemblies.³⁵⁻³⁸ *Cyclotella cryptica* was published in 2016 and was previously available, but has since been removed from the UCSC Genome Browser where it had previously been deposited.³⁷ Two other diatom projects (*Fragillariopsis cylindrus* and *Pseudo-nitzschia multiseries*) are in draft form.³⁸⁻

⁴⁰ *P. tricornutum* and *T. pseudonana* are the most advanced and characterized diatom genomes available for study.^{35, 36}

With only two diatom genomes (*Phaeodactylum tricornutum* and *Thalassiosira pseudonana*) that are assembled to the chromosome level and several other diatom genomes in draft, RGd-1 represents a novel genome from an extremophilic, alkaline, freshwater stream in Yellowstone National Park (YNP). Analysis of the RGd-1 genome will provide insight into novel genes and specialized pathways that may allow RGd-1 to accumulate high lipid concentrations.

Dissertation Overview

To date there have been 2 marine diatoms that have had their genomes sequenced and made publicly available. The RGd-1 genome is markedly different from any other diatom sequenced because it is an extremophilic, freshwater diatom. Furthermore, with long-range PacBio reads, RGd-1 was found to have a potential bacterial symbiont, *Brevundimonas sp.* Using Illumina® Mi-Seq sequencing it was possible to obtain a community analysis of bacteria associated with RGd-1 in culture. A total of 9 prokaryotic operational taxonomic units (OTUs) were identified and further research could elucidate their roles as potential phycosphere bacteria, where they may have specific functional roles with RGd-1 in the area around an algal cell.⁴¹⁻⁴⁸

Following the introduction in Chapter One, Chapter Two titled “Biodiesel (Microalgae)” from the book, “Extremophilic Microbial Processing of Lignocellulosic Feedstocks to Biofuels, Value-Added Products, and Usable Power”, provides an overview of extremophilic algae, isolation, and their use for biodiesel production and secondary high-value coproducts.

Transitioning from a broad overview of extremophilic algae for use in biofuels, Chapter Three discusses how to isolate strains, determine whether they are unialgal using sequencing methods and determine which strains are promising for further optimization through physiological characterization. Specifically, 11 strains that were isolated from Witch Creek, Yellowstone National Park, WY, USA, and two culture-collection strains were grown with and without sodium bicarbonate addition. To facilitate the analysis, the strains were divided into fastest growers, lipid producers, and biomass producers to determine if they were promising for biodiesel production or other applications.

After examining algae on a physiological-level, Chapter Four discusses the RGd-1 genome sequencing, assembly and annotation. This organism was chosen because it was found to naturally contain 30-40% (w/w) triacylglycerol and 70-80% (w/w) fatty acid methyl esters that can be transesterified into biodiesel,⁴⁹ among the highest biofuel potential (BP) published in the literature. The majority of this chapter discusses the RGd-1 genome and 16S amplified results for the microbial community in the RGd-1 culture that is thought to be the phycosphere. Of particular focus is the *Brevundimonas* sp. genome assembly that was assembled as part of an RGd-1 Pac-Bio sequencing project. Collectively, a genome assembly and its associated community members are referred to as the hologenome, or the entire genome.⁵⁰ With the RGd-1 genome assembly and data from the 16S amplified sequencing, it is possible to gain insight into the RGd-1 ecological interactions, ecology, and evolution.

The *Brevundimonas* sp. genome was introduced in Chapter 4, and Chapter Five builds on this by providing the genome announcement for the *Brevundimonas* sp., in preparation for publication including its genome assembly statistics, and genes of interest that were identified

from the genome annotation. Finally, Chapter Six summarizes the work performed in this dissertation and proposes future directions.

13
CHAPTER TWO

BIODIESEL (MICROALGAE)

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Karen Moll

Contributions: Wrote the book chapter

Co-Author: Todd Pederson

Contributions: Wrote the book chapter

Co-Author: Robert D. Gardner

Contributions: Wrote the book chapter

Co-Author: Brent M. Peyton

Contributions: Discussed, commented, and edited the book chapter

Manuscript Information

Karen Moll, Todd Pederson, Robert D. Gardner, Brent M. Peyton

“Extremophilic Microbial Processing of Lignocellulosic Feedstocks to Biofuels, Value-Added Products, and Usable Power.” Chapter 4, “Biodiesel (Microalgae), D.R. Sani, Editor. 2018, pp 63-78, Springer: New York, NY.”

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☒ Published in a peer-reviewed journal

Algal biomass represents a promising renewable energy system due to fast photoautotrophic growth rates, CO₂ fixation, and accumulation of carbon storage metabolites which can be used as precursors for fuels and specialty chemicals; moreover, it offers a solution to offset the global dependence on conventional fuels. Currently, transportation fuels make up approximately 66% of the global energy demand.^{51, 52} Correspondingly, extensive use of non-renewable energy sources has increased the global carbon dioxide (CO₂) concentration approximately 43% since the use of these fuels was significantly intensified since the Industrial Revolution, around 1750 (USEPA 2016). Fossil fuels are subject to volatile price swings based on geopolitical issues and availability of crude oil. As sources of crude oil become depleted, prices associated with fuels and other petroleum-derived products will experience rapid increases in response. Biofuels and other bio-products derived from microalgae have potential to contribute significantly to this market, yet how large their impact on the market will be remains to be seen.⁵²

Microalgae are oxygenic, phototrophic eukaryotes, which are abundantly found across diverse environments ranging from acidic hot springs to arctic ice and snow. Like other phototrophs, microalgae require light energy, water and a few inorganic nutrients (carbon dioxide, nitrogen, phosphorus, iron, etc.) which they convert to biomass with diverse biochemical composition.⁵³ There are several advantages to utilizing microalgae for biofuel production. First, microalgae have increased theoretical photosynthetic efficiency (10–12%) over terrestrial plants (4-6%) and high cell division rates (1–3 day⁻¹) which leads to overall improved biomass yield per unit area, which when paired with the ability to be cultivated continuously year-round further improves their productivity over terrestrial plants.⁵⁴ Additionally, microalgae

can be grown using brackish, salt, and wastewater sources reducing their demand for freshwater, and can utilize nitrogen and phosphorus from agricultural, industrial, and municipal wastewaters as low-cost nutrient sources and as a method for remediation of the wastewater.⁵⁵ Furthermore, algae have the potential to reduce carbon emissions if co-located with a power plant to sequester portions of the emitted CO₂ before it enters the atmosphere.⁵⁶ Lastly, and perhaps most importantly, microalgae frequently have higher lipid content than terrestrial plants,⁵⁷ and with the combination of the other added benefits, often result in higher biofuel productivities on a per biomass basis. The inherent advantages to using microalgae for sustainable biofuel and bio-product formation are well known, though several bottlenecks still exist on the path from lab bench to full-scale production of microalgae.

One of the primary challenges associated with scale-up of biofuel production is algal species selection. For rapid growth in large-scale open systems, a robust species that tolerates moderate temperature, pH and salinity changes must be selected to keep productivity high. Extremophilic algae are a compelling choice for biofuel production because of their innate ability to survive and even thrive on the boundary of extreme conditions.⁵⁸ Extremophilic microalgal strains have an added benefit of growing in conditions that inhibit growth of many competing microorganisms, which may allow higher biofuel productivity of targeted strains. Some strains isolated from alkaline or halophilic environments have been shown to contain very high concentrations of lipids, primarily in the form of triacylglycerol (TAG). Further, alkaline environments have greater flux of atmospheric CO₂ into the algal growth medium, thus increasing inorganic carbon available for fixation. Therefore, extremophilic, microalgal strains have the potential to improve algal biofuel viability by providing a more cost-effective

production with a greater potential for algal biodiesel productivity and decreased probability for significant contamination.

Microalgae

Algal biofuels are derived from two predominant groups, green algae and diatoms, both of which are unicellular, photosynthetic eukaryotes (Figure 2.1). Some strains are known to store high concentrations of triacylglycerol (TAG) that can be converted into biodiesel. Diatom strain, RGd-1, was found to produce 30–40% (w/w) TAG and 70–80% (w/w) biofuel potential (BP) for ash-free dry weight.⁴⁹ An isolate from the Heart Lake area of Yellowstone National Park, RGd-1 is able to grow in exceptionally high silica concentrations that are often inhibiting for marine diatoms.⁵⁹ Moll et al. (2014) found that RGd-1 maintained the best growth and TAG accumulation when grown in 2 mM Si, which is roughly an order of magnitude greater than the silica concentration in seawater.⁴⁹ They went on to further stimulate TAG accumulation by adding 25 mM NaHCO_3 just prior to nutrient depletion. The greatest lipid accumulation occurred during the stress of a combined Si and NO_3^- limitation with NaHCO_3 addition which yielded a nearly a two-fold increase in TAG accumulation compared to Si limitation alone (Figure 2). Further, NaHCO_3 addition increased TAG accumulation compared to only nutrient limitation.⁴⁹

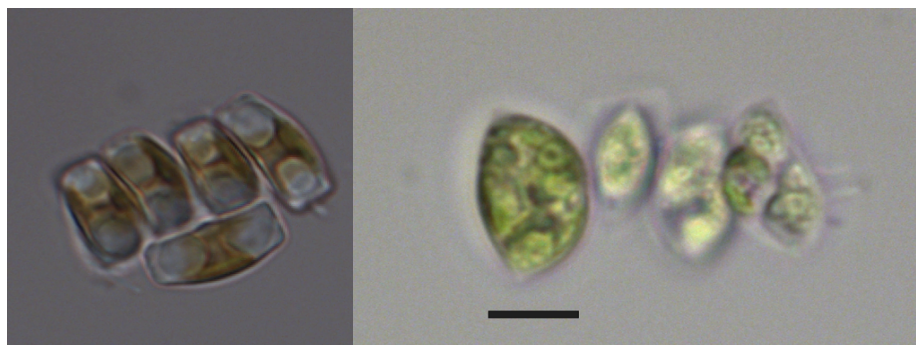


Figure 2.1 YNP diatom strain RGd-1 (left) and YNP green algal WC-1 (right, scale bar 10 μm).

Chlorophytes (green algae) are thought to utilize the C_3 photosynthesis pathway for carbon fixation, whereas diatoms, including *Thalassiosira pseudonana*, are thought to use the C_3 and C_4 pathways.^{36, 60, 61} However, both mechanisms utilize ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo) to catalyze the first step in the Calvin cycle. RuBisCo has a relatively low affinity for CO_2 and is less than half saturated under normal atmospheric conditions.⁶² By consequence, microalgae have evolved carbon concentrating mechanisms (CCMs) to increase the carbon flux to RuBisCo.⁶³⁻⁶⁵ Algal CCMs are essential two phase processes. In the first phase, inorganic carbon is acquired from the environment and shuttled to the chloroplast. During the second phase, HCO_3^- is increased in the chloroplast stroma.⁶⁵

Microalgae have a number of carbonic anhydrases and bicarbonate transport channels to move inorganic carbon across the periplasmic membrane, through the cytosol, into the chloroplast, and convert the carbon to CO_2 in the direct vicinity of the Rubisco in the pyrenoid.

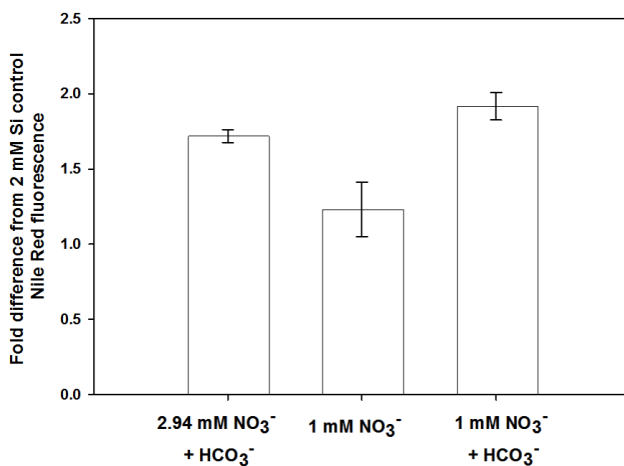


Figure 2.2 Each bar represents the fold difference in Nile Red fluorescence intensities at 15 d for each treatment compared to the 2 mM Si control.

Interestingly, C₄ pathways have the extra ability to convert HCO₃⁻ directly to a C₄ organic acid molecule which is shuttled to the pyrenoid and reconverted to CO₂ to be used by RuBisCo.^{61, 66}

Alkaline environments (e.g., Soap Lake, Washington) often have high bicarbonate ion concentrations. Given the current understanding of CCM's it is not surprising that soda lakes are highly photosynthetically productive. Organisms that thrive in these extreme environments have physiological adaptations that allow them to be successful under conditions that would be lethal to other microorganisms. Diatoms are uniquely suited to living in alkaline environments due to their C₄ metabolism. Evidence from the *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* genomes indicates a propensity for C₄ metabolism. Valenzuela et al. (2012) found evidence for *P. tricornutum* using C₃ and C₄ metabolism when dissolved inorganic carbon concentrations were low.⁶⁷ Further, they found an increase in expression for *P. tricornutum* pyruvate carboxylase, malic enzyme and malate dehydrogenase which indicates the presence of the C₄ pathway.⁶⁷ This is advantageous by providing another pathway for CO₂ fixation, especially given that C₄ carboxylases are high affinity molecules allowing carbon to be concentrated in the chloroplast. As more extremophilic microalgae are isolated, identified and characterized, further advances in biotechnology for biofuels and renewable biochemicals will become available.

Under replete growth conditions, microalgae capable of TAG accumulation will synthesize TAG during light hours and utilize the stored carbon for cellular maintenance during dark hours.⁶⁸⁻⁷⁰ However, when the cellular cycling is stressed or arrested due to nutrient limitation environmental stress (e.g., pH, light, temperature stress),⁷¹⁻⁷⁴ or by chemical addition,^{75, 76} many algal strains will accumulate and maintain TAG vacuoles within the cell. Thus, industrial algal biofuel systems producing TAG as a biofuel substrate have to balance rapid growth with a means

of impeding the cell cycle when the culture has reached a desired density.^{77, 78} Typically, this is accomplished by timing cellular density with the depletion of nitrogen in the growth medium, however this can often make the culture susceptible to contamination or predation from other microorganisms. However, use of extremophilic strains as an industrial algal biofuel platform is an under studied tactic and merits additional investigation focused on cellular cycling and TAG accumulation.

Extreme Environments

Microalgae have been isolated from extreme environments such as Arctic/Antarctic regions and acidic hot springs, as well as from alkaline and/or hypersaline environments. Examples of such environments include Yellowstone National Park, Wyoming; Soap Lake, Washington; Mono Lake, California; Great Salt Lake, Utah; and the East African Soda Lakes. In addition to providing a selective advantage for algal growth, with increased pH, there is increased CO₂ solubility, leading to enhanced algal growth.⁷⁹ Soda lakes accumulate very high concentrations of sodium carbonate salts due to the limited Mg²⁺ and Ca²⁺ concentrations with pH 8-12.⁸⁰ The East African Soda Lakes are among the most productive lakes in the world with gross photosynthetic rates up to 36 g O₂ m² day⁻¹ for Lake Nakuru, Kenya.⁸¹ Another example of a soda lake from which high lipid containing strains have been isolated is Soap Lake, Washington which is pH 9.9 and contains very high concentrations of sodium carbonate and sodium bicarbonate at 6870 mg L⁻¹ (0.7%) and 5209 mg L⁻¹ (0.5%), respectively.⁸² Halophiles are uniquely adapted to their environments by keeping high concentrations of intracellular K⁺ to compensate for the high extracellular Na⁺ concentrations. Pick et al. (1986) found that when

Dunaliella salina was grown in 1-4 M NaCl, intracellular Na⁺ concentrations were 20-100 mM and K⁺ concentrations were 150-250 mM.⁸³

Witch Creek is an alkaline, freshwater creek located in the Heart Lake area of Yellowstone National Park (WY, USA). Witch Creek is approximately two miles long and is fed by a combination of fresh groundwater and effluent channels (Figure 2.3) from alkaline hot springs with high concentrations of metals such as arsenic (~300 ppb) and silicon (~72 ppm). Regular inputs from these thermal features into Witch Creek make the creek alkaline, leading to the growth of microorganisms including microalgae that are adapted to the alkaline conditions found in Witch Creek. Such microalgae have been isolated and characterized for biofuel applications.



Figure 2.3 Inputs from thermal hot springs into Witch Creek (research in Yellowstone was conducted under an approved Yellowstone Research Permit [Permit # 5480]).

Targeting Extremophiles

For microalgae, extremophilic organisms are those considered to have improved growth outside of “normal” environments. For microalgae, these defined “normal” environmental parameters are outlined as in Seckbach (2007) as those having an optimum temperature range

between 4–40°C, a pH range of 5–8.5, and a salinity range between that of fresh and salt water (0%–3.5%). The bulk of extremophilic microalgae species fall into the alkaliphile, acidophile, or halophile classifications although there are thermophilic microalgae as well.⁵⁸ Although the majority of microalgae species find their optimum growth somewhere in these defined ranges, extremophilic species have been targeted for use in biofuel production because of the generalized acceptance that product formation of lipids and other products is increased when the cell cycle is ceased and environmental stresses are implemented.^{15, 84, 85} Table 2.1 highlights several extremophilic algae including acidophiles, alkaliphiles, psychrophiles, thermophiles and halophiles, and the conditions from which they were isolated.

Table 2.1 Examples of extremophilic microalgae and their desirable temperature, pH, and salinity conditions with each having at least one environmental condition outside of normal range.

Extremophilic Organism	Extremophilic Condition			Reference
	Temp	pH	Salinity	
<i>Dunaliella salina</i>	0-38°C	6-9	3-31%	Borowitzka (1990)
<i>Cyanidium caldarium</i>	35-55°C	2-3	<3%	Doemel and Brock (1971)
Yellowstone Diatom Isolate Rgd-1	28°C	9.3	<3%	Moll et al. (2014)
Yellowstone Green Isolate WC-1 <i>Scenedesmus</i> sp.	24°C	9.3	<3%	Gardner et al. (2010)
<i>Chlamydomonas nivalis</i>	1.5-20°	6-8	<3%	Remias et al. (2005)
<i>Aphanothece halophytica</i>	30°C	7.5	15-30%	Madigan et al. (2008)
<i>Cyanodioschyzon</i> sp.	45-55°C	2.5	<3%	Skorupa et al. (2014)

Bioprospecting

Bioprospecting for potential strains that can be used for biodiesel production begins by matching desired growth conditions (*e.g.*, high pH and salinity values) with natural environments that contain those conditions. In addition, locations for microbiological sampling are either on public or private property, and written permission to collect samples should be obtained for any samples collected. For locating extremophilic microalgae, find environments with a pH value below 5 or above pH 8.5, and a salinity range above that of salt water (3.5% w/w) up to sodium chloride saturation (35% [w/w]). For temperature limits, the upper limit for microalgal (eukaryotic) growth is approximately 57°C.⁵⁸ However, around 45°C phototrophic growth may be dominated by cyanobacterial species. Therefore, with regards to temperature, microalgal growth will be primarily in the mesophilic range (20-45°C).⁸⁶ As shown in Figure 4, microbial and microalgal communities can change significantly over very short distances due to gradients in temperature, pH, salinity, or nutrient availability, therefore care should be taken to characterize specific sampling locations for optimum selection of targeted microorganisms. Once areas have been targeted for sampling and written permission is obtained, samples can be taken from the area of interest and returned to the lab for isolation.



Figure 2.4 Typical heterogeneity of a sampling site containing green algae, diatoms and cyanobacteria (research in Yellowstone was conducted under an approved Yellowstone Research Permit [Permit # 5480]).

In the approach recommended here, samples should be disaggregated and inoculated into 5 mL of various microalgal media types (e.g., Bold's Basal Medium) to determine which would provide the best conditions for growth. Typically, standard algal growth media are not ideal for isolation of extremophilic microalgae, and so must be adjusted to higher or lower pH, higher salinity, or temperature must be controlled at a higher or lower value. Samples should also be streaked for isolation on the appropriate solid growth media. Once the colonies have grown to sufficient size, individual, isolated colonies should be aseptically "picked" from the agar plate and inoculated into liquid medium (1 mL) and grown until they change the color of the medium (often green or brown, for green algae or diatoms, respectively), and subsequently transferred to liquid medium (5 mL). After approximately two weeks or once they have reached substantial growth, cultures should be streaked for isolation again and re-picked from agar plates for a total of three rounds of streaking for isolation and transferring to new medium to ensure isolation from other algal species. Following each round of isolation, strains should be observed under transmitted light microscopy to determine the cellular morphology.

Each isolated strain should be screened for TAG content by staining with Nile Red dye or Bodipy 505-515 and observed using epifluorescent microscopy.⁸⁷ Fluorescence of lipid vacuoles appear bright yellow for Nile Red dye under epifluorescent light (Figure 2.5) or green when stained with Bodipy. Strains that have the highest TAG content should be selected for further characterization, since these will likely have the highest biodiesel potential. Strains found to have high concentrations of TAG during the screening process should also have their growth rates measured, since a combination of fast growth rate and high lipid production is desirable. Strains that show faster growth rates and TAG production should be inoculated into approximately 150

mL of liquid medium in 250 mL baffled flasks and grown in triplicate at various temperatures to determine the optimal growth and TAG accumulation conditions for each strain.

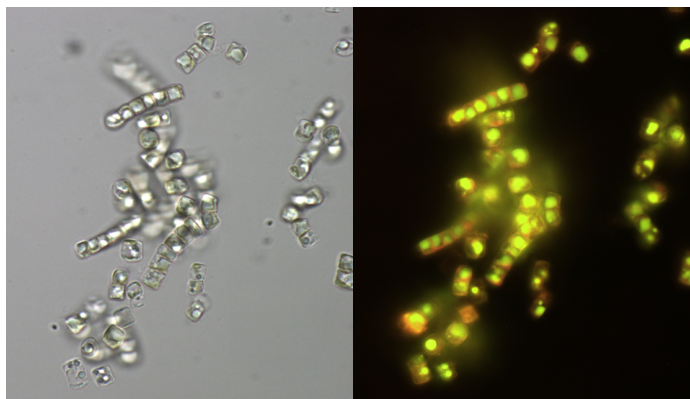


Figure 2.5 RGD-1 transmitted light (left) and Nile Red fluorescence under epifluorescent light (right).

Algae as Biofuels

Manufacturing of biofuels and bioproducts from microalgae at industrial scale has many proposed methodologies for start-to-finish generation of targeted end products. Similar to a conventional petroleum refinery which makes multiple products and fuels from crude petroleum, a biorefinery would produce biofuels and other products from algal biomass. An operation such as this would have a variety of different steps, but the major sequences are as follows: cultivation of algae for biomass generation, harvesting of algal biomass, and extraction/conversion of algal biomass. Significant life cycle analyses and techno-economic analyses will be needed with regard to each step of this process to determine the most advantageous strategy for each phase of the operation.⁸⁸

Primary cultivation systems for the production of algal biomass are through the use of raceway ponds (Figure 2.6) or photobioreactors (Figure 2.7). Raceway ponds are typically large

closed-loop ponds, in which microalgal culture is continuously circulated through a designed path, generally with the use of revolving paddle wheels.⁵ Algae are most often grown in large outdoor raceway ponds because they are one of the most cost efficient ways to grow large quantities of algae;⁵⁶ however, raceway ponds can suffer from large evaporative losses and poor mass transfer properties for the application of CO₂.⁸⁹ Still, the major caveat of this type of system is that it is non-sterile and has a potential for undesirable contamination from faster growing microorganisms that are not biofuel productive. One advantage of using extremophilic algae in outdoor raceway ponds is to create pond conditions that are favorable for extremophilic algal growth while being inhibitory for faster growing microorganisms that do not produce biodiesel precursors, but are relatively low-cost to build and operate. Conversely, photobioreactors are collections of small to medium diameter (< 10 cm) transparent tubes which are all oriented parallel to each other and usually stacked vertically to increase the reactor volume in a given footprint.⁵ Primary concerns regarding the use of photobioreactors are the design limitations on tube length, which is dependent on the degree of O₂ production, CO₂ depletion, and pH variation.⁵⁴ While photobioreactors can cost more to build, they typically offer a more controlled environment and higher productivity than open raceway ponds.³



Figure 2.6 Outdoor raceway pond (2000L) at Utah State University, Logan UT.

Even dense cultures of microalgae require removal of excess water for downstream processing to some extent, and cost-effective and energy-efficient processes for the removal of water and subsequent harvesting of algal biomass are required for economical production of algal biofuels.⁵⁶ Harvesting of the algae is an energy intensive step because many conversion pathways require the algal biomass to be at substantially higher concentrations than cultures grow in nature. Typically, even for extremely productive strains, biomass concentrations will not exceed 5% [w/w] suspensions while most conversion strategies require a minimum 20% biomass slurry and can require even more dewatering and drying. There are many different approaches to harvesting of algae, but the major developed strategies for harvesting algae are



Figure 2.7 Photobioreactor illuminated (Green Wave Energy, Inc.) by artificial light in pilot-scale laboratory setting at Montana State University, Bozeman MT.

flocculation and sedimentation, filtration, and centrifugation. Each of these methods has advantages and disadvantages, and these methods are often used in combination to reach the desired final algae to water ratio. Flocculation and sedimentation is a routine method for harvesting algae which do not settle out in well-maintained reactors because of their small cell size.⁵¹ Flocculation can be obtained through chemical additives such as alum, lime, polyacrylamide polymers or surfactants. Following flocculation of the cells, the cells are allowed to settle, and excess water can be removed from the top of the cell sediment. Flocculation is also commonly used with dissolved air flotation (DAF) where the flocculated biomass is driven upward by the attachment of microbubbles where it can be collected at high concentration at the tank surface. Another common form of harvesting microalgae is centrifugation. It is likely that

centrifugation will play a minor role in harvesting of culture where other harvesting methods fail to reach the desired algae content for slurry, as centrifugation can reach higher concentration biomass slurries than flocculation with sedimentation or other harvesting alternatives. However, centrifugation is an energy intensive process which makes it seemingly unappealing for scale-up of algae cultivation for biofuels.⁹⁰ Filtration is another possible alternative for harvesting algae, but has lost some appeal in scale-up from laboratory testing because of the potential for membrane or screen fouling as well as the labor-intensive process of operating such a system. Combinations of these practices for harvesting algae can reach biomass concentrations in the 20%–30% [w/w] range required for conversion methods utilizing wet biomass, but are not ideal for methods that require whole or dry biomass. If further dewatering is needed after the culture harvesting, a drying step will be necessary for removing excess water from algae paste or slurry. Thermal drying using methane drum dryers is most commonly practiced, but other oven-type dryers have been used, as well as solar drying and freeze-drying of algae slurry.⁵¹

Conversion of algal biomass can be accomplished through different methods which are generally categorized into two categories, thermochemical conversion or biochemical conversion.⁵⁷ From cultivated microalgal biomass there are two major conversion strategies for making usable biofuels. The first, and more well known, is biochemical conversion. Most common for microalgal biodiesel production is the process of transesterification. Through the use of heat and an acid or base catalyst, algal lipids are converted to fatty acid methyl esters (FAME's) and glycerol. These FAME's are crude biodiesel and are similar in composition to those produced from transesterification of vegetable oils. The high lipid content in microalgae, primary TAG, makes transesterification an efficient process with production yields between 70%–90%. While transesterification is primarily a straightforward chemical reaction, it falls into

the biochemical conversion category because it does not require the significant energy requirements for high temperature and pressure systems typical with thermochemical conversion pathways. Another biochemical conversion pathway is fermentation of an algal slurry to produce ethanol. Fermentation of algae is a less common method for energy production, mostly because of the difficulties associated with the process of separating produced ethanol after the fermentation process as well as the relatively low starch content of microalgae compared with alternative lignocellulosic biomass. Still, fermentation of lipid extracted algae for conversion of the residual carbohydrates may offset costs associated with biofuel production from microalgae.

As an alternative to biochemical conversion, thermochemical conversion is currently being heavily studied for its application to microalgal biomass. Of the many different techniques for thermochemical conversion, hydrothermal liquefaction and pyrolysis are emerging as the two benchmark technologies, while gasification and hydrogenation will likely play smaller roles in utilizing all products from the conversion process.⁹¹ Hydrothermal liquefaction is a process which uses sub-critical water at moderate temperature and pressure (~300 °C and 10 Mpa) to convert wet biomass into a liquid fuel called primary oil. This oil can be separated and purified using a solvent such as dichloromethane. Other products from hydrothermal liquefaction such as the aqueous and gas phases can be recycled to supply nutrients for more algae cultivation or used in a gasification or hydrogenation process for other products. Pyrolysis is another thermochemical conversion process to produce energy rich compounds such as biofuel, charcoal and gaseous products from algal biomass. Short residence times and high temperatures (500°C) are used to crack biomass into short chain molecules which can then be rapidly cooled into a liquid phase to produce biofuel. Pyrolysis has a high-energy input required because it requires algal biomass to be completely dried, adding the necessity for 20%–30% algal slurries to have all

remaining residual water removed through one of the processes mentioned previously. There is still no general consensus on whether biochemical or thermochemical conversion processes will be the ultimate solution to producing biofuel from microalgae; however, life cycle analysis and techno-economic analysis considering the entire production of algae to biofuels is being completed with considerations for both types of conversion methods.⁹²

Other Secondary Products

Microalgae not only offer a source of sustainable biodiesel, but can be used to make an assortment of products such as food supplements, fertilizers, bioplastics, nutraceuticals and cosmetics.¹⁵ For algal biofuels to be economically viable, additional co-products will need to be formed in concert with biofuel. In particular, those products which have a combination of the highest yields and highest specific selling price (e.g., \$/lb.) will be the optimum targets for coproducing with biofuels. Two examples of these higher value compounds are carotenoids and unsaturated fatty acids, which are both produced naturally by microalgae. Carotenoids are colorful pigments which can be used as food and feed additives, as well as nutraceutical supplements to promote health. The two most common carotenoids produced naturally by microalgae are β -carotene and astaxanthin¹⁵. Two specific organisms produce these compounds in much greater quantity than any others studied, *Dunaliella salina* and *Haematococcus pluvialis* for β -carotene and astaxanthin production, respectively. Coincidentally, *Dunaliella salina* is a halophile with optimum sodium chloride concentrations between 10%–27% depending on targeted growth regime, but also produces culture rich in β -carotene. Alternatively, *Haematococcus pluvialis* produces high concentrations of astaxanthin when environmentally stressed with sodium chloride concentrations in the range of 4%–6%. Furthermore, both these

organisms can withstand high light environments and even can be induced to make more of their targeted carotenoids with light induced stress.¹⁵

Mono- and poly-unsaturated fatty acids being another alternative secondary product, they must be valued separately from being simply a biofuel precursor. Depending on the intended application, omega-3 fatty acids such as α -linolenic acid or eicosapentaenoic acid can be sold as health supplements for much higher value as opposed to being converted to biofuel. Similarly, the monounsaturated fatty acid oleic acid is a valuable precursor to 9-decenoic acid, which can be used to create valuable products such as surfactants, lubricants, and polyester, amongst other things.⁹³ A feasible production strategy would make not only biofuels, but some other valuable secondary products as well.

Take Home Message

- Microalgae are promising candidates for biodiesel production due to their fast growth rates, can be cultivated on non-arable land, can use brackish or wastewater and avoid competing with food supplies.
- Some microalgae strains accumulate high concentrations of TAG that can be converted into biodiesel.
- Extremophilic algae are uniquely suited for growth because their growth conditions inhibit growth for most contaminating (biofuel non-productive) microorganisms.

33
CHAPTER THREE

CHARACTERIZATION OF NINE NOVEL GREEN
ALGAE STRAINS FROM YELLOWSTONE
NATIONAL PARK

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Karen Moll

Contributions: Designed and performed experiments and primary writer

Co-Author: Robert D. Gardner

Contributions: Designed experiments, performed initial experiments

Co-Author: Todd Pederson

Contributions: Discussed and performed experiments, contributed to writing

Co-Author: Muneeb S. Rathore

Contributions: Discussed and performed experiments, contributed to writing

Co-Author: Brent M. Peyton

Contributions: Discussed, commented, and edited the manuscript

Manuscript Information

Karen M. Moll, Robert D. Gardner, Todd Pederson, Muneeb S. Rathore, Robin Gerlach, Brent M. Peyton

Algal Research

Status of Manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☐ Published in a peer-reviewed journal

35
Abstract

Many algae have been isolated and characterized for biofuel production capabilities. This screening study evaluated nine green algae strains isolated from an alkaline stream in Yellowstone National Park (YNP). Two highly characterized culture-collection strains, PC-3 and UTEX395, *Coelastrella* sp. and *Chlorella* sp. were characterized in parallel. Here, we describe the methods used for strain isolation and evaluation from field samples to bench-scale strain characterization for potential biomass and biofuel production. Each strain was evaluated for growth rate, biomass, and lipid production. The 11 strains were grown with and without sodium bicarbonate addition to determine whether added inorganic carbon could stimulate an increase in lipid accumulation. While no single strain was ideal for all three evaluation criteria, several strains were promising for two of the three criteria indicating that those strains might be good candidates for biodiesel or biomass production. Strain WC-5 had a fast growth rate and high fatty acid methyl ester (FAME) content and was considered the best candidate for biodiesel production. WC-2b performed the best for biomass production due to a fast growth rate and comparatively high final dry-cell weight (DCW) compared to the other strains.

Keywords: Algae, Alkaliphile, Biodiesel, Biofuel, Extremophile, TAG, Yellowstone

36
Introduction

Petroleum-based fuels are not sustainable as a long-term resource for transportation fuels. As a renewable energy source, algal biofuels can be used to alleviate transport-fuel demands.⁹⁴ With the exception of recent economic impacts of the Covid-19 pandemic, the largest spikes in fuel prices have triggered economic crises for 10 of 12 U.S. recessions since World War II.^{95,96} While it may appear that reserves are increasing, oil that can be accessed most easily is diminishing.^{95, 96} This has resulted in drilling for oil in more dangerous and less ideal locations with higher environmental impacts; with exploration of drilling into environmentally, culturally,⁹⁷ historically, or archaeologically sensitive locations such as Arctic National Wildlife Refuge, Standing Rock, Bears Ears, and Grand Staircase Escalante. As of 2016, there were 13 National Parks with energy production within park borders,⁹⁸ and an average of 550 active oil and gas wells are within the U.S. National Park System each year.⁹⁹

A potential alternative to petroleum-based transportation fuels is algal biofuel. The benefits of algal biofuels over second-generation (ethanol) biofuels are multiple: algae require less land area for growth, have faster doubling times than cellulosic crops, and do not increase the prices of food-stocks.⁹⁴ To improve algal biofuel viability, strain selection has been identified as a key factor for improvement.^{3, 77, 94, 100} Fast growth rates, high lipid accumulation and/or high biomass concentration are the most important considerations with regard to algal biofuel success. Fast growth rates increase the probability of successful biofuel production using algal monocultures or defined mixed cultures. When algal growth rates are fast, there is an increased potential that an algal pond will be productive for biofuel by outcompeting contaminating (micro)organisms, or low lipid-yielding algae strains.¹⁰¹ Open systems, such as outdoor raceway ponds, are susceptible to unwanted algal contamination by wind, rain, animals, and other non-

sterile sources.¹⁰² With slower growth rates, algae may permit faster growing organisms such as bacteria and other common contaminants, such as golden algae,¹⁰³ cyanobacteria, protists and bacteria that are naturally found in water, to consume limiting nutrients required by algae for growth and lipid accumulation.¹⁰³⁻¹⁰⁵ If nutrients such as nitrate, phosphate, sulfate, silica, iron or magnesium have been depleted by non-productive organisms, conditions for a “pond-crash” will result as defined by McBride et al. (2014) and Carney et al. (2016).¹⁰⁴⁻¹⁰⁶

Some algae strains naturally accumulate high concentrations of lipids in the form of triacylglycerol (TAG) that can be converted to biodiesel. Previous studies have shown that some green algae and diatoms can increase the accumulation of TAG and other lipids by approximately two-fold when sodium bicarbonate is added at the appropriate time.^{49, 70, 107, 108} However, not all strains respond to this stimulation and each strain must be evaluated to determine whether sodium bicarbonate addition will increase lipid accumulation.

While some strains accumulate lipids, others are known to reach high cell densities and dry cell weight (DCW) with low lipid concentrations. The ability for these strains to quickly divide and reach high cell densities diverts fixed carbon into new biomass or cells rather than into long-chain fatty acids, such as TAGs. Strains that have fast growth rates and accumulate high concentrations of TAGs or reach high DCW are the most promising strains for commercial production of desirable algal products. Strain evaluation and improvement are therefore of the utmost importance.

Promising algae strains have been isolated from a variety of unique environments including extreme environments such as the polar regions (Arctic/Antarctic), alkaline lakes such as Soap Lake, Washington, East African Soda Lakes (Lake Nakuru, Kenya), Great Salt Lake, Mono Lake, California and YNP.^{69, 71, 80, 81, 101, 109-114} Targeting algae from these extreme

environments may have added benefits. For example, alkaline growth conditions may inhibit the growth of many contaminating microorganisms as well as provide increased CO₂ availability.^{3, 41, 49, 101} Furthermore, the high pH can potentially pose an added “stress” to further induce lipid accumulation.^{49, 107, 115}

Here, we provide an in-depth analysis of nine green-algae strains that were isolated from YNP, WY, USA, and two reference culture-collection strains for growth rate, biomass, and biodiesel potential (BP). The nine alkaline-adapted YNP strains and two culture-collection strains were evaluated for their potential for biodiesel or biomass production.

Materials & Methods

Samples collected from Witch Creek in Yellowstone National Park, WY, USA (pH of 9.3),¹¹⁶ were streaked for isolation on agar plates prepared from the media types; Bold’s Basal Medium adjusted to pH 8.7 (B8.7), ASP-2 200, Bristol’s Medium, ASP-2-Fresh, and B8.7SiS. The ASP-2-based media types contained approximately twenty-times the NO₃⁻ concentrations (59 mM NaNO₃) compared to Bristol’s and Bold’s Basal Medium-based media types (2.94 mM NaNO₃).^{49, 117-120} Cultures were grown in direct sunlight or in a light incubator (Percival) for at least one week. Isolated green colonies were selected from the streak-plates and inoculated into 1.0 mL of sterile liquid medium, incubated for one to two weeks, and streaked for isolation again. A total of three rounds of streaking for isolation and picking individual colonies into 1.0 mL of sterile liquid medium were performed. For each round of isolation, colony and cellular morphologies were documented. At the time of each transfer, colonies were exposed to Nile Red fluorescent stain (Sigma Aldrich) to assess the lipid accumulation within the cells⁸⁷ and were

examined with light microscopy to examine cell morphologies. Strains that contained relatively large lipid vacuoles were selected for additional characterization (Figure 3.1).

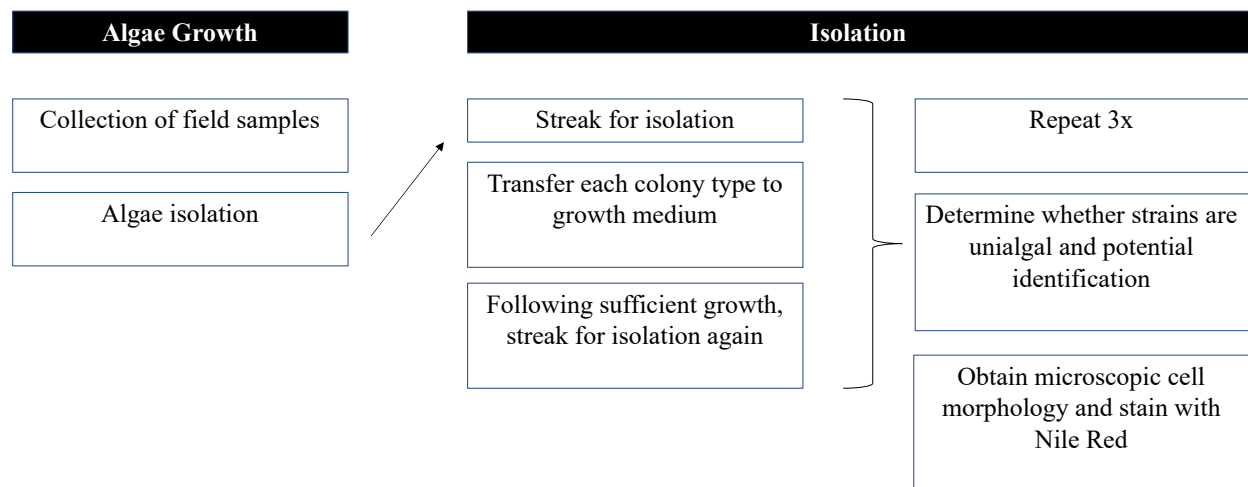


Figure 3.1 Outline for algae isolation beginning from field collection to strain characterization. Each strain was streaked for isolation on solid growth medium, grown in liquid growth medium and visualized microscopically at each step to ensure strain isolation.

Screening Studies

Following the three rounds of isolation, the culture volume was increased from 1 mL to 5 mL to 75 mL to 150 mL to ultimately obtain a culture volume sufficient for experimental testing. For growth studies, cultures were inoculated in triplicate flasks in B8.7⁴⁹ and grown at room temperature ($25^{\circ}\text{C} \pm 2^{\circ}\text{C}$) with light provided at $400 \mu\text{mole photons m}^{-2}\text{s}^{-1}$ using twelve T5 four-foot fluorescent lights in a square-wave 14:10 light/dark (L/D) cycle. Experimental parameters that were measured include: cell concentrations which were monitored using a hemacytometer (Reichert), pH (Accumet) and Nile Red fluorescence described in detail below. Cultures were grown for at least 10 days or until Nile Red fluorescence showed a peak indicating that cultures had reached a maximum lipid accumulation.⁴⁹ The cultures exhibiting the fastest growth rates

(lowest doubling times) and highest Nile Red fluorescence were selected for additional in-depth characterization.

In-depth Characterization Studies

At this stage, eleven cultures (9 YNP isolates and 2 culture collection strains) were each grown in triplicate B8.7 medium in 1L photobioreactor (PBR) tubes, in a temperature-controlled circulating-water bath ($25 \pm 2^\circ\text{C}$).^{49, 115, 117} Each strain was inoculated into two sets of triplicates. One triplicate received sodium bicarbonate (NaHCO_3) addition and was compared to the other set of triplicate cultures that served as an air-only control. Prior to sodium bicarbonate addition, the replicates are referred to as replicates 1 and 2, after which, they are called air-only or sodium bicarbonate condition. Cultures were sparged with ambient air at a flow rate of 0.4 L min^{-1} , which contained approximately 400 ppb of atmospheric CO_2 or $8\text{--}9 \text{ mg C L}^{-1}$ dissolved inorganic carbon (DIC) in the growth medium, as measured by flow meters and carbon analyzer measurements, respectively.

Bicarbonate Addition

To determine if sodium bicarbonate addition increased TAG accumulation, 25 mM sodium bicarbonate was added before nitrogen depletion at the beginning of the 14:10 L:D light cycle.¹⁰⁷ The nitrate concentration in the medium was monitored using the szechrome NAS assay to ensure sufficient nitrate was available prior to sodium bicarbonate addition.^{107, 121} Dissolved inorganic carbon was measured using a Skalar Formacs TOC/TN Carbon Analyzer with an LAS-160 autosampler (Skalar). The DIC concentration was determined through injection of $100 \mu\text{L}$ of sample in 2% (v/v) phosphoric acid; CO_2 peak area was measured with an infrared (IR) detector and calibrated with peak responses from sodium carbonate/bicarbonate standard solutions.⁴⁹

Determination of Unialgal Strains and Strain Identification

DNA was extracted from fifty mL of each strain and sequenced using 454 Pyrosequencing to determine that they were unialgal. DNA was extracted using a CTAB/bead beating method outlined by Moll (2014)⁴⁹ and quantified using a Qubit fluorometer with a dsDNA BR Assay kit. The 454-Pyrosequencing was performed according to Bowen de Leon et al., 2012, and Bell et al., 2018.^{41, 122, 123} Reads were analyzed using Qiime¹²⁴ and redundant reads were clustered and eliminated with CDHIT-EST.¹²⁵ The phylogenetic tree was constructed using RaxML (version 8.2.12) and FigTree (version 1.4.4).^{126, 127} To obtain resolved identifications, each strain was amplified across the internal transcribed spacer region ITS 1 (ITS 1: 5' TCCGTAGGTGAACCTGCGG 3'), 5.8S, and ITS 2 (ITS 4: 5' TCCTCCGCTTATTGATATGC 3') regions with Sanger Sequencing (Table 3.2).¹²⁸⁻¹³¹ Each culture was periodically verified as axenic when grown on solid medium (B8.7 with 0.5% glucose, 0.5% yeast extract and 2% agarose) at room temperature ($25 \pm 2^\circ\text{C}$) in the dark.

Dry Cell Weight

Twenty five mL of suspended culture were filtered GF/F Glass Microfiber Filters (Whatman) to determine the DCW. To remove water, each sample was lyophilized (Labconco Lyophilizer). All samples were weighed immediately upon retrieval from the lyophilizer to minimize air vapor condensation.⁴⁹

Nitrate

To ensure sufficient nitrate was available in the growth medium, the colorimetric szechrome NAS assay was used prior to sodium bicarbonate addition.^{107, 121} The Dionex™ ICS-1100 Ion Chromatography System (IC) was also used to verify and obtain a more accurate analysis of the anions (including nitrate) in each sample. The ion chromatography system consisted of a Dionex™ IonPac AS22 separation column coupled with a Dionex™ AERS 500 (4 mm) suppressor. Nitrate peaks were reported in μS using a Thermo Scientific DS6 Heated Conductivity Cell. The peaks were integrated and compared against the standard calibration curve made using the Dionex™ Combined Seven Anion Standard I solution. The samples were diluted (1:5) when necessary using nanopore water (MilliQ) to stay within the linear range of quantification.

Nile Red Fluorescence

To obtain a daily estimation of TAGs for each culture condition, samples were stained with Nile Red (9-diethylamino-5H-benzo (α) phenoxa-zine-5-one) (Sigma- Aldrich).^{49, 69, 72, 87, 115} Samples were diluted 1:5 using ultra-filtered water and suspended in 20% DMSO or acetone (Table A.1).¹³² Each sample was stained with 20 μL of Nile Red (Sigma-Aldrich). The optimal stain times^{49, 71, 107, 133} had been determined for each strain prior to experimentation ranging from 4 to 60 min where stained green algae cells were quantified on a microplate reader (BioTek, Synergy H1) with 530/575 (20% DMSO) and 480/580 (acetone) excitation/emission filters. The time after staining when a maximum peak in Nile Red fluorescence relative fluorescence units (rfu) was observed indicated the optimal time at which the stain permeated the cell wall and stained neutral lipids in the lipid vacuoles.⁴⁹

Results

Each of the nine YNP strains was verified to be unialgal by 454-Pyrosequencing and analysis by Qiime before the strain characterization studies.¹²⁴ To facilitate analysis, the strains were first categorized by the three evaluation criteria: growth rate, lipid yield, and biomass production. The strains within each of those categories were then characterized in further depth for each of the parameters tested.

Verification of Unialgal Strains and Strain Identification

Previously, algae strains were primarily determined to be unialgal by cell morphology when viewed microscopically. Today, methods such as next generation sequencing (NGS), enable more robust determination of culture purity. Each of the nine YNP isolates was confirmed unialgal by 454 Pyrosequencing and analysis with Qiime¹²⁴ before the characterization studies. The resulting extracted DNA concentrations from the YNP green algae strains ranged from 30.1 to 497 ng/ μ L for PGV10-G2 and WC-5, respectively (Table 3.1).

Table 3.1 SSU rDNA (18S) DNA concentrations for the extracted, amplified, and purified prior to 454-Pyrosequencing for the nine YNP green algae strains. DNA was quantified using a Qubit fluorometer with a dsDNA BR Assay kit.

Strain	Extracted DNA (ng/ μ L)	DNA Concentration (ng/ μ L) after PCR cleanup for 454-Pyrosequencing
MF-1	115	67.0
PGV-6	140	78.9
PGV8-G1	43.4	63.4
PGV8-G2	73.1	48.6
PGV10-G1	237	52.1
PGV10-G2	30.1	48.2
WC-1	252	56.8
WC-2b	157	45.1
WC-5	497	57.6

Strain purity was determined using the approximately 500 bp amplified V1-V3 region of 18S SSU rDNA sequenced by 454 Pyrosequencing. Strain identity was determined using the full-length ITS1, 5.8S and ITS2 regions by Sanger sequencing (Tables 3.2 and 3.3, Figure 3.2).

Table 3.2 Representative sequences for 18S SSU rDNA. Each sequence was BLAST searched for identification. There were three strains that had identical BLAST results with different identifications for 18S (MF1, PGV6, and PGV10-G1), which represents the diverse collection of sequences in NCBI.

Algae isolate	Identification	% ID	Query Coverage	E-value	Accession	Amplicon length (bp)
MF1	<i>Chlorobion braunii</i>	95.73%	100%	0.0	gi 933801142 KT833591.1 gi 19847825 X91263.1	513
	<i>Podohedriella falcate</i>	95.73%	100%	0.0		
PGV6	<i>Tetradasmus obliquus</i>	95.05%	100%	0.0	gi 1341122479 MG022741.1 gi 930306211 KR082492.1	525
	<i>Scenedesmus acutus</i>	95.05%	100%	0.0		
PGV8-G1	<i>Chlorococcum sp.</i>	93.74%	99%	0.0	gi 563323151 KF791553.1	523
PGV8-G2	<i>Chlorococcum sp.</i>	94.41%	100%	0.0	gi 563323151 KF791553.1	482
PGV10-G1	<i>Desmodesmus sp.</i>	98.45%	87%	0.0	gi 693012472 AB917137.1 gi 402704215 JX258841.1	518
	<i>Scenedesmus sp.</i>	98.45%	87%	0.0		
PGV10-G2	<i>Chlorella pyrenoidosa</i>	96.77%	85%	0.0	gi 5326955 AJ242762.1	544
WC-1	<i>Tetradasmus obliquus</i>	97.56%	74%	0.0	gi 1345677876 MG971386.1 gi 1245685193 KY816917.1	551
	<i>Scenedesmus sp.</i>	97.56%	74%	0.0		
WC-2b	<i>Chlorococcum sp.</i>	99.27%	100%	5e-136	gi 563323151 KF791553.1	498
WC-5	<i>Chlorella sorokiniana</i>	94.94%	100%	0.0	gi 1176441302 KY921855.1	514

Table 3.3 BLAST identification of ITS amplicons obtained by Sanger sequencing. The sequence identity was determined by BLAST search for identification. There were two strains that had identical BLAST results with different identifications ITS (PGV6 and WC-1).

Sample	Identification	% ID	Query Coverage	E-value	Accession	Amplicon length (bp)
MF1	<i>Scenedesmus sp.</i>	99.32%	100%	0.0	gi 530330741 KF471115.1	1162
PGV6	<i>Acutodesmus obliquus</i>	100%	100%	0.0	gi 998524167 KU170646.1 gi 359385305 FR865721.1	1224
	<i>Scenedesmus obliquus</i>	100%	100%	0.0		
PGV8-G1	<i>Chloromonas sp.</i>	98.26%	25%	9e-77	gi 312270216 HQ404890.1	1358
PGV8-G2	<i>Chloromonas sp.</i>	98.26%	28%	9e-77	gi 312270216 HQ404890.1	1194
PGV10-G1	<i>Desmodesmus sp.</i>	99.65%	100%	0.0	gi 1520100899 MH010842.1	1132
PGV10-G2	<i>Desmodesmus sp.</i>	100%	100%	0.0	gi 1569048762 MK496927.1	1119
WC-1	<i>Tetradesmus obliquus</i>	99.53%	71%	0.0	gi 1341122479 MG022741.1 gi 1273809524 MF326554.1 gi 1027901709 KU291882.1	1752
	<i>Acutodesmus sp.</i>	99.53%	71%	0.0		
	<i>Scenedesmus basilliensis</i>	99.53%	71%	0.0		
WC-2b	<i>Botryococcus sp.</i>	93.79%	99%	0.0	<u>gi 558605714 KF537773.1</u>	1177
WC-5	<i>Chlorella sp.</i>	95.73%	99%	0.0	gi 1321358367 MG757661.1	1298

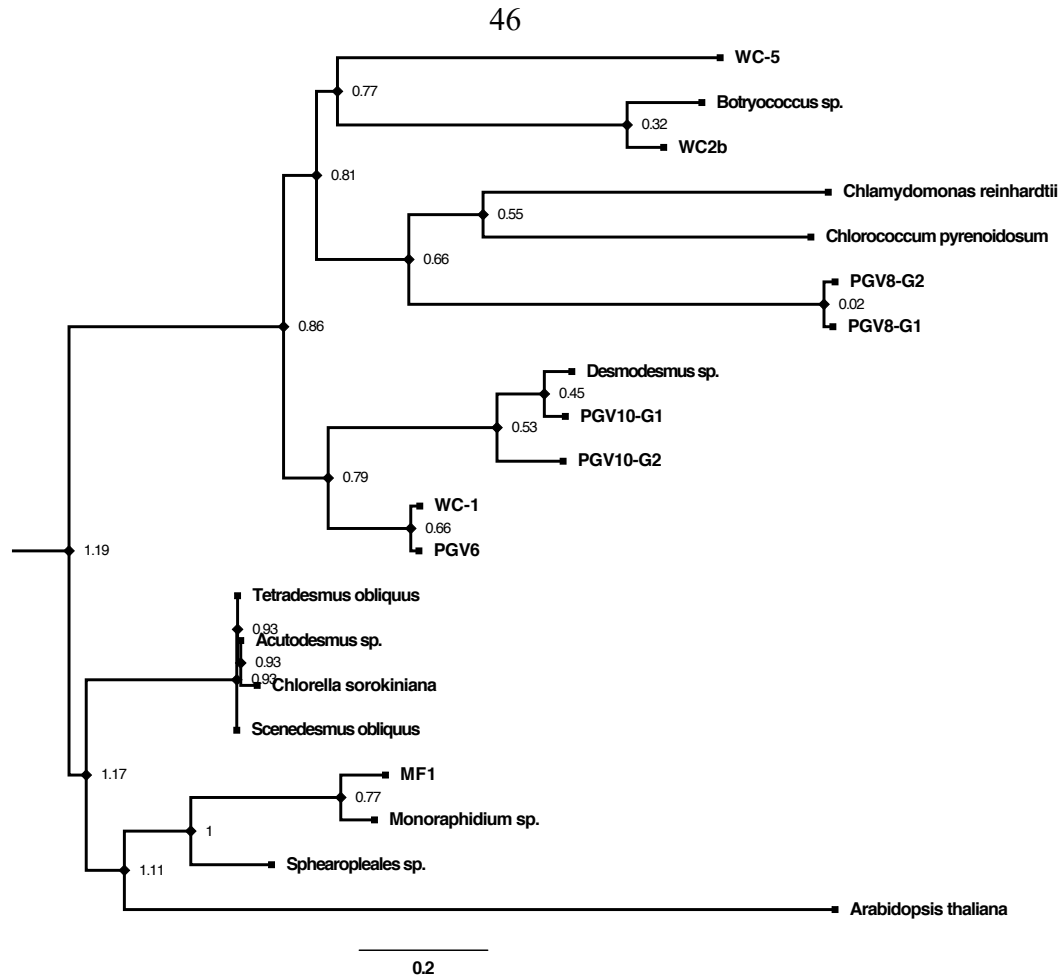


Figure 3.2 Full-length ITS Sanger results used for strain identification. Each strain (median = 1194 bp), was aligned with Muscle (v3.8.1551)134 and phylogenetic distances were determined using the Maximum Likelihood method with RaxML (version 8.2.12).127 The scale bar represents number of nucleotide changes between strains. The bootstrap values present at the nodes represent the divergence event on a time scale.

Doubling Time

Three of eleven strains exhibited distinctly faster doubling times compared to the others (Figure 3.3). The strains exhibiting the fastest doubling times were WC-2b, WC-5, and WC-1 and at 15.98 ± 1.48 h (air-only), 15.00 ± 0.50 h (replicate 2) and 16.07 ± 1.45 h (replicate 2) (average $\pm$ 95% confidence interval), respectively. The slowest growing strains were PGV6, PC-

3, and UTEX395 with the doubling times 30.40 ± 2.84 h (air-only) and 27.45 ± 1.45 h (sodium bicarbonate addition) (Figure 3.3).

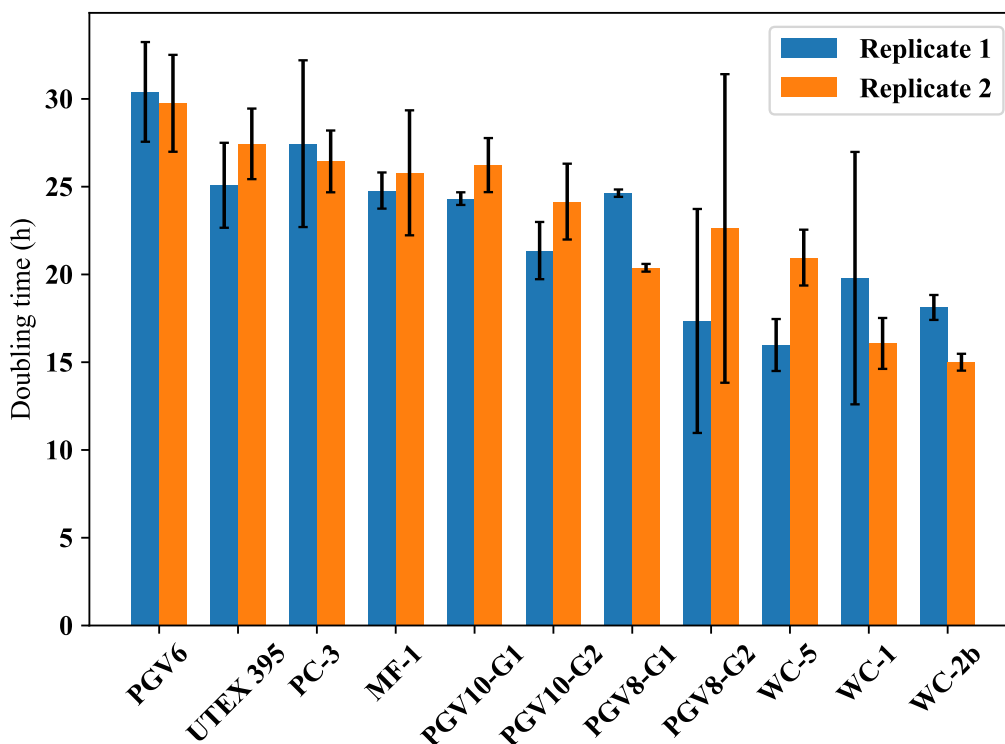


Figure 3.3 Average doubling times based on the maximal growth rates for each of the eleven strains (replicate 1 and replicate 2). The error bars represent 95% confidence intervals.

While there was a slight increase in doubling time in replicate 2 for each of the three strains, the air-only and sodium bicarbonate conditions were within error of each other. Figure 3.4A shows that WC-5 reached the highest final cell concentrations ($7.75 \times 10^7 \pm 6.34 \times 10^6$ cell mL^{-1} and $7.66 \times 10^7 \pm 8.64 \times 10^6$ cell mL^{-1} for the air only and sodium bicarbonate conditions, respectively) on the day of harvest, day 13.5. On day two there was a large amount of error in cell counts due to the nature of the WC-2b cell division, which was highly variable from PBR to PBR. WC-2b has a colonial cellular organization, such that when a mature cell lyses, it can

release multiple smaller cells.¹³⁴ In asynchronous cultures, the cell lysis will be offset, causing large variation in less dense cell cultures.

Among the fastest-growing strains, WC-5 reached the highest pH. The pH of the air only and sodium bicarbonate conditions were similar at 10.88 ± 0.56 and 10.41 ± 0.41 , respectively (Figure 3.4B). The pH was greater than 10.5 indicating that the culture was likely carbon limited and inorganic carbon in the culture primarily existed as carbonate rather than bicarbonate (Figure 3.4B). There was a noticeable pH decrease for each culture just prior to the sodium bicarbonate additions. This was due to the accumulation of dissolved CO₂ during the dark-period, just before the bicarbonate additions were made.

As shown in Figure 3.4C, each strain depleted the medium of NO₃⁻ between days 6 and 10. The starting nitrate concentration in Bold's is 182.4 mg L⁻¹, but after the photobioreactors were autoclaved, some evaporative loss occurred causing the nitrate concentration to vary slightly. The medium NO₃⁻ concentrations were measured just prior to sodium bicarbonate addition to ensure there were sufficient concentrations available. Previous results have shown that the addition of sodium bicarbonate after NO₃⁻ depletion results in a drastic decrease in lipid concentrations.¹¹⁵ Nitrate was determined to be available for the three fastest growing strains, ensuring that lipid accumulation could result following sodium bicarbonate addition. As shown in Figure 3.4C, each strain depleted the medium of NO₃⁻ between day 6 and 10.

The Nile Red fluorescence showed an increase after nitrate depletion (Figure 3.4C, D). The addition of sodium bicarbonate changed the Nile Red fluorescence intensity of WC-5, WC-1 and WC-2b, with an increase of 45%, -13%, and 37% respectively, compared to the air only condition (Figure 3.4D, Table Appendices A2). Strain WC-5 reached the highest Nile Red

fluorescence with nearly one order of magnitude in difference for both conditions compared to WC-1 and WC-2b.

Biomass Production

As shown in Figure 3.5, the final highest DCW occurred after sodium bicarbonate addition, for nine of eleven strains. At harvest, the strains with the highest DCWs were PC-3 and WC-2b at 1.50 ± 0.14 and 1.44 ± 0.10 g L⁻¹, respectively, for the sodium bicarbonate condition. The strains with the lowest final DCWs were PGV10-G1, PGV8-G1 and PGV8-G2 at 0.69 ± 0.10 (sodium bicarbonate addition), 0.56 ± 0.11 (air only), and 0.54 ± 0.10 (sodium bicarbonate addition) g L⁻¹, respectively (Table 3.3). The strains that were unaffected by sodium bicarbonate addition were PGV-6, WC-5, PGV10-G1, PGV10-G2 and PGV8-G2 (Figure 3.5).

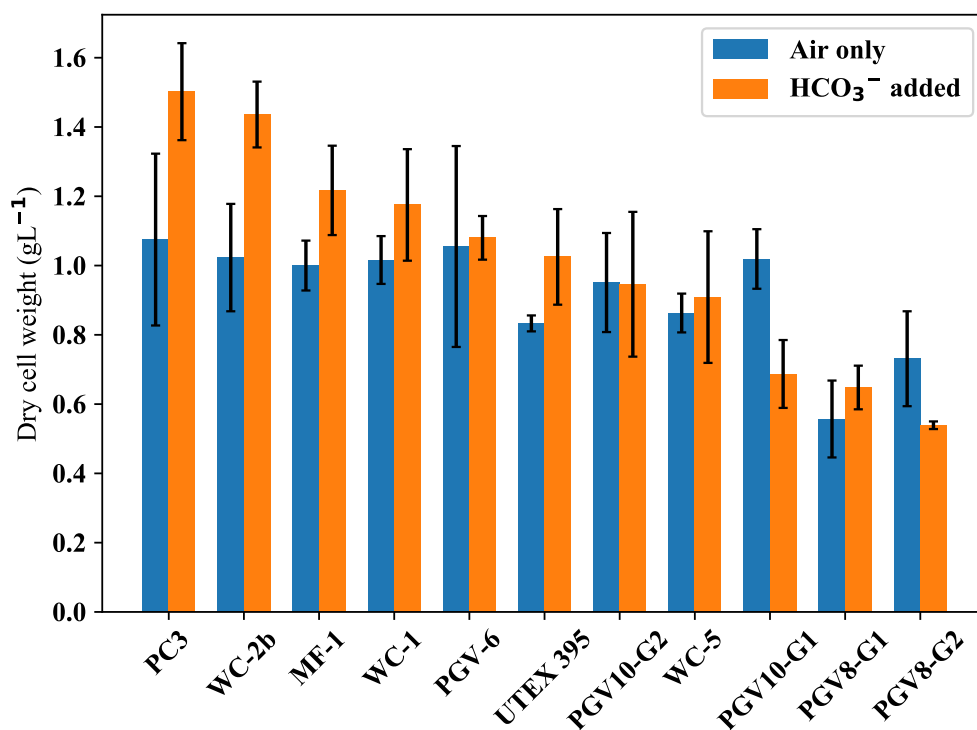


Figure 3.5 Average DCW at harvest for each of the eleven strains (with and without sodium bicarbonate addition). The error bars represent 95% confidence intervals.

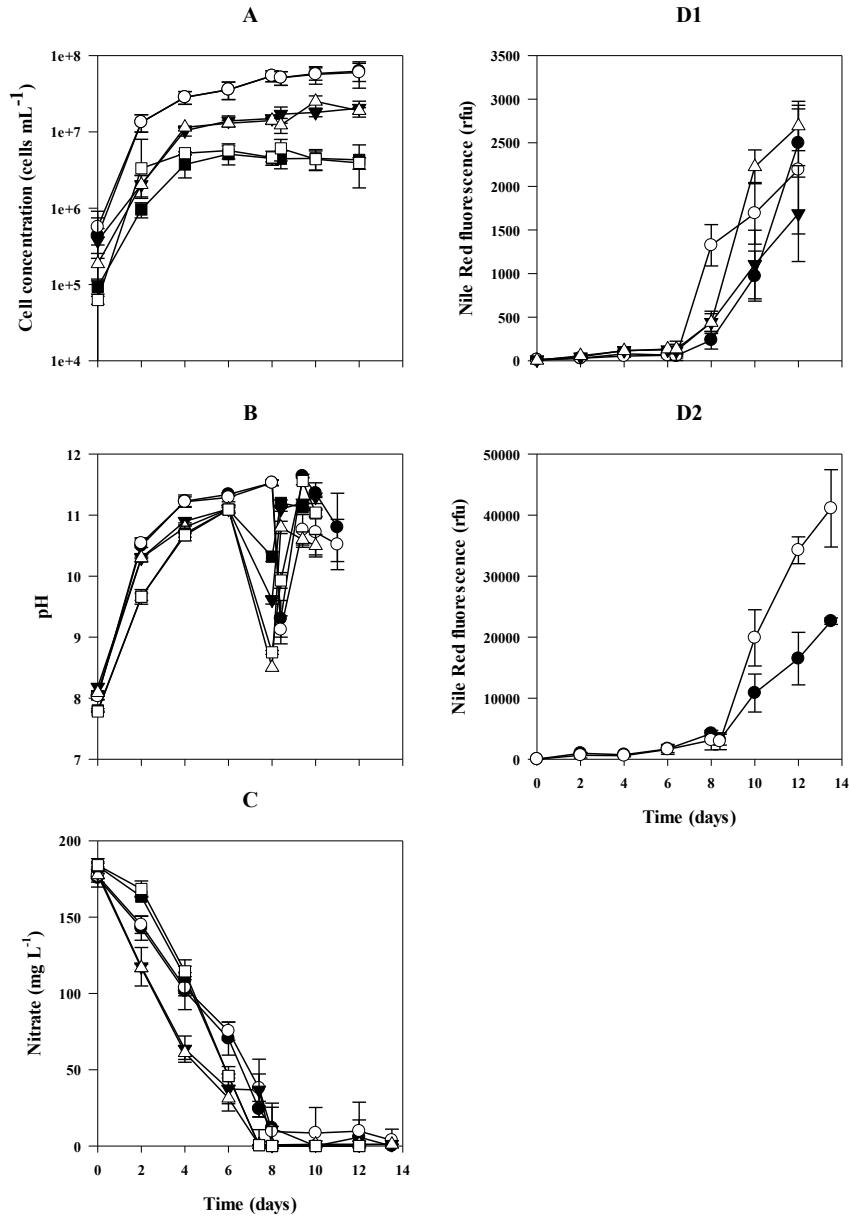


Figure 3.4 Growth and lipid accumulation observations for the cultures with the fastest growth rates, WC-1, WC-2b and WC-5 with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C) and Nile Red fluorescence (D) for the fastest doubling times. The error bars represent 95% confidence intervals. For Figures A-C, the series are represented by the following: WC-5 (circles), WC-1 (triangles) and WC-2b (squares). In Figure D1, WC-1 (circles) and WC-2b (triangles) and Figure D2 WC-5 (circles). The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.

PC-3 and WC-2b were the two strains that produced the highest endpoint dry cell weight 1.08 ± 0.25 and 1.50 ± 0.14 and 1.02 ± 0.15 and 1.44 ± 0.10 , respectively, for air-only and sodium bicarbonate addition (Figure 3.5). While WC-2b had faster doubling times than PC-3 (Figure 3.3) (18.13 ± 0.71 h and 27.45 ± 4.75 h), PC-3 had a higher maximum cell count by stationary phase of growth at $1.04 \times 10^7 \pm 5.18 \times 10^6$ cell mL⁻¹ compared to WC-2b at $6.08 \times 10^6 \pm 1.91 \times 10^6$ cell mL⁻¹ (Figure 3.6A) indicating higher biomass production.

At pH values greater than 10.3 the majority of the carbonate species will primarily exist as CO₃²⁻, which is difficult to impossible for some algae to fix.⁷⁰ At high pH values, some algae strains experience a “pH stress” which may contribute to lipid accumulation.⁴⁹ Once sodium bicarbonate was added, the pH ranged from 9.86 – 10.26 during the light periods, due to the buffering effect of sodium bicarbonate (pKa = 10.3).¹³⁵ WC-2b reached the highest pH of the two high biomass producers at 11.56 ± 0.06 at day 8 for the sodium bicarbonate condition. When grown in the air-only condition, PC-3 maintained a higher pH at approximately 11.10 compared to the sodium bicarbonate condition that ranged from 9.86 – 10.38, due to its buffering ability (Figure 3.6B).

Once nitrate was depleted (Figure 3.6C), the Nile Red fluorescence increased on days 8 and 10 for WC-2b and PC-3, respectively (Figure 3.6D). PC-3 had a maximum Nile Red fluorescence of 1553 ± 363 and 1678 ± 166 relative fluorescence units (rfu) for the air-only and sodium bicarbonate addition conditions, respectively (Figure 3.6D). WC-2b had a maximum Nile Red fluorescence of 1688 ± 489 and 2693 ± 250 rfu. While there was not a significant increase in Nile Red fluorescence when sodium bicarbonate was added to the PC-3 cultures, there was a 37% increase in fluorescence for WC-2b, indicating a greater concentration of neutral lipids.

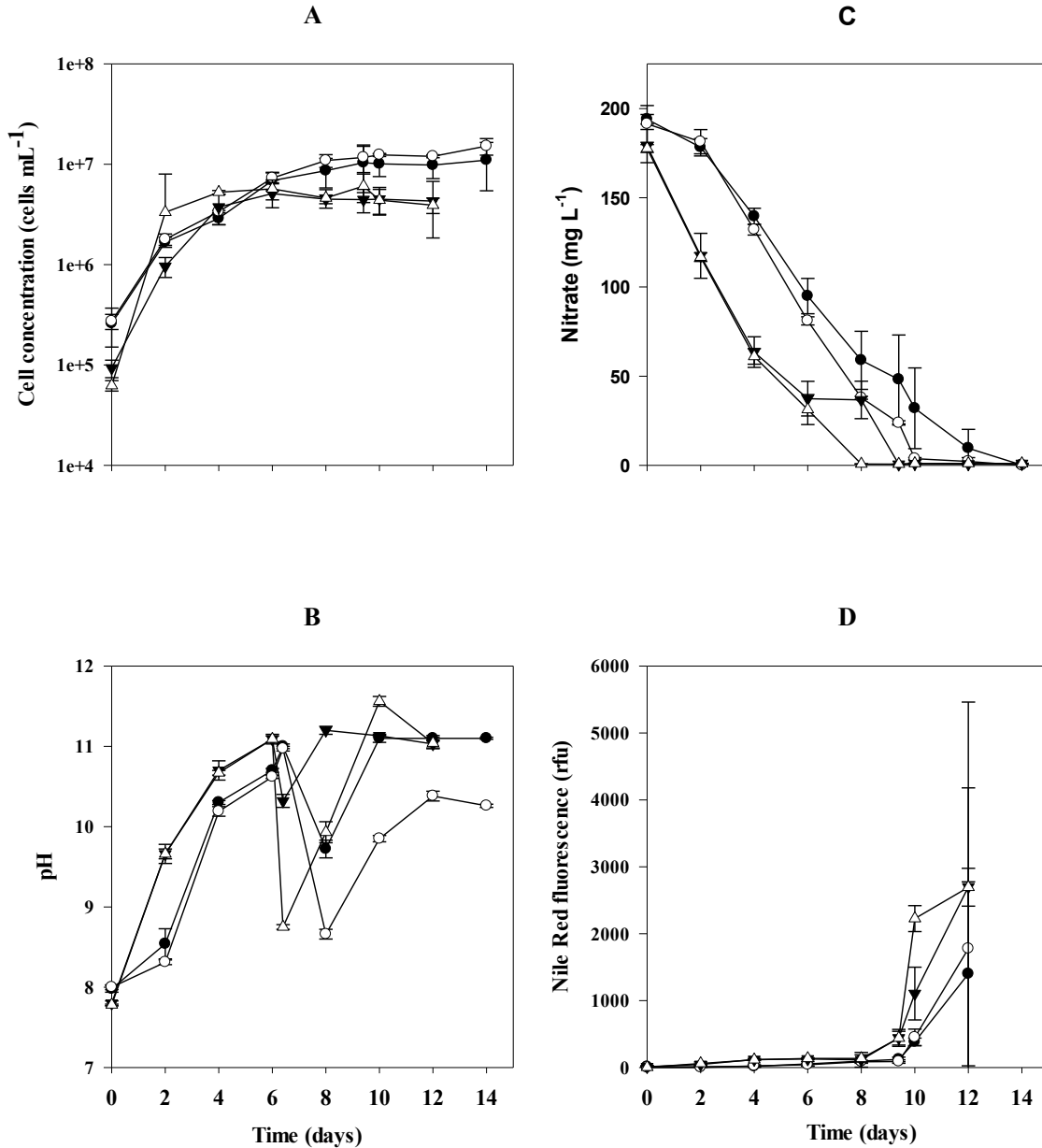


Figure 3.5 Growth and lipid accumulation observations for the cultures with the highest biomass production as DCW, PC-3 (circles), and WC-2b (triangles), with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C) and neutral lipids (D). The error bars represent 95% confidence intervals. The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.

Highest Lipid Producing Strains

Using Nile Red screening, the two highest lipid producing strains were WC-5 and UTEX395. Compared to the next highest strain in Nile Red fluorescence, MF-1 at 18928 ± 2549 rfu, WC-5 and UTEX-395 were approximately double in fluorescence at 41120 ± 6342 , and 34960 ± 1522 rfu, respectively. The two strains that reached the highest DCW, PC-3 and WC-2b, were the lowest in Nile Red fluorescence.

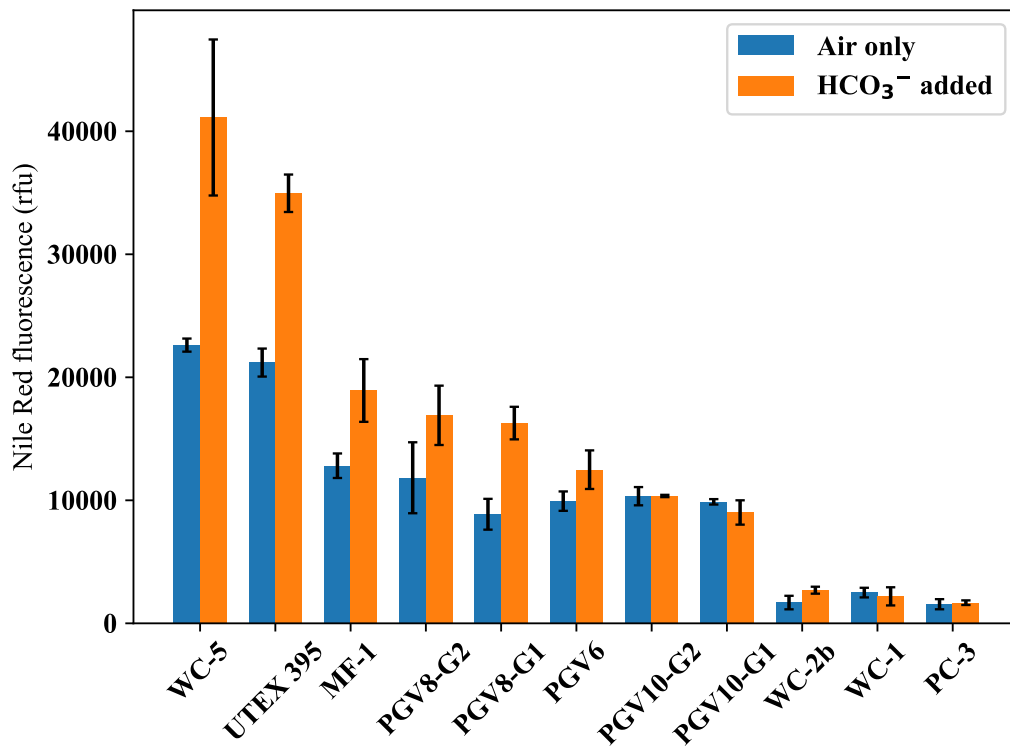


Figure 3.6 Average final Nile Red fluorescence for each of the eleven strains (with and without sodium bicarbonate addition). The error bars represent 95% confidence intervals.

Among these strains, WC-5 grew the fastest at a doubling time of 16.00 ± 1.48 h for the air only condition compared to UTEX-395 that doubled at a rate of 25.08 ± 2.42 h. (Figure 3.8A). However, both strains were similar in their final cell concentrations with WC-5 reaching

$7.75 \times 10^7 \pm 6.34 \times 10^6$ cell mL⁻¹ and UTEX-395 at $7.66 \times 10^7 \pm 8.64 \times 10^6$ cell mL⁻¹. Sodium bicarbonate was added at the beginning of the light cycle following a 10-hour dark period. Because photosynthesis did not occur during this time, dissolved CO₂ in the growth medium accumulated in the growth medium which can be observed by a drop in pH at 6.4, 7.4, and 8.4 days (Figure 3.8B). The lower pH at the beginning of the light cycle compared to the normal sampling time at the end of the light cycle is due to the accumulation of H₂CO₃ in the growth medium.¹¹⁵ Following this measurement, sodium bicarbonate was added. Nitrate was verified to be present in each of the samples just before sodium bicarbonate addition (Figure 3.8C) on day 7.4.

Among the highest lipid producers, the strain that reached the highest Nile Red fluorescence intensity was WC-5 (Figure 3.8D). With the addition of sodium bicarbonate, UTEX 395 showed a 65% increase in Nile Red fluorescence above the air-only condition. Among the best lipid producing strains, the Nile Red fluorescence of 4213.33 ± 495.36 (replicate 1) 3116.67 ± 1580.67 (replicate 2) for WC-5 and 11263.33 ± 763.83 (replicate 1) 11036.67 ± 436.26 (replicate 2) for UTEX 395 started to increase on days 6 and 7.4, respectively, which coincided with a pH stress and nitrate limitation.

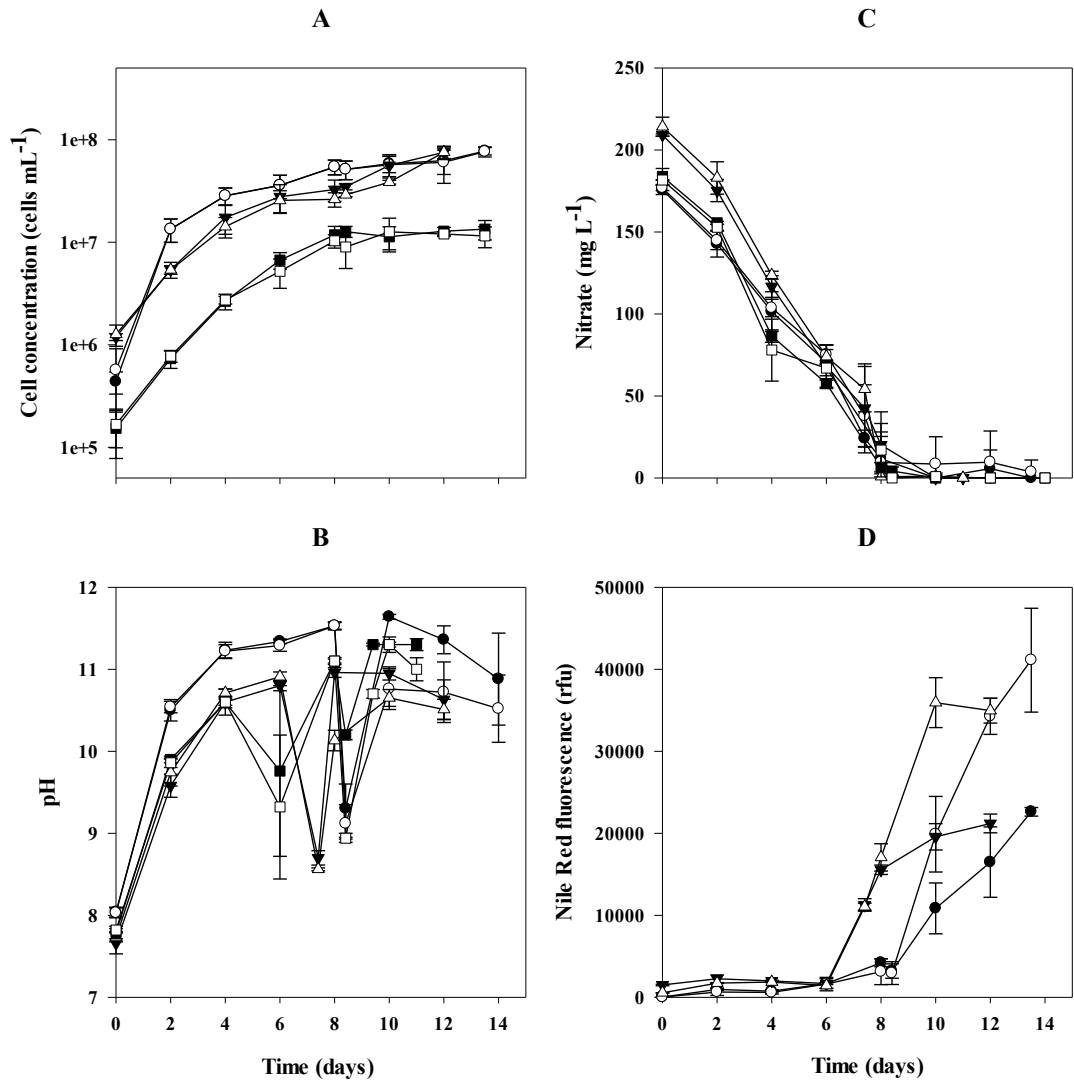


Figure 3.7 Growth and lipid accumulating observations for the cultures with the highest lipid production, WC-5 (circles) and UTEX 395 (triangles), with and without sodium bicarbonate addition: cell concentration (A), pH (B), medium nitrate concentration (C), and FAMEs (D). The error bars represent 95% confidence intervals. The open and filled symbols represent air only and sodium bicarbonate addition, respectively, for all conditions.

Discussion

As more strains are isolated, potentially novel properties may be identified such as lipid accumulation abilities in specialized environments and high-value coproducts. Eleven green algae strains were screened for their potential use in biodiesel production. Strains were grouped based on three properties: fastest growth, highest biomass yield, and highest lipid production.

Growth rate

Of the eleven strains evaluated, the three fastest growing strains were WC-5 (*Chlorella*), WC-1 (*Tetradesmus*), and WC-2b (*Botryococcus*). Additionally, some *Botryococcus* sp. are known to accumulate high lipid concentrations,¹³⁶⁻¹³⁸ while WC-2b was one of the lowest strains in Nile Red fluorescence. Fast growing strains are more likely to be productive in outdoor raceway ponds because they can outcompete other non-algae strains for nutrients in the growth medium. Fast growing strains may also be productive for biomass and or lipid concentrations. Identification of high biomass producers and lipid producers can determine which strains should be used for certain applications, such as wastewater remediation or biodiesel production.

Biomass production

Based on doubling time and DCW, the best strain for biomass production characterized here was PC-3. This strain may be promising in applications where biomass production is combined with nutrient removal in wastewater treatment and bioremediation of NO_3^- and PO_4^{3-} .¹³⁹⁻¹⁴³ Eustance et al. (2013) indicated that *Scenedesmus obliquus* and *Monoraphidium* sp. were able to utilize ammonia, nitrate or urea while concomitantly producing biodiesel for two green algae strains and could potentially be used in wastewater remediation.^{144 145-147} Further,

microalgae have been studied for remediation of nutrient-rich swine and cattle waste or municipal wastewater while eliminating ammonia and phosphate and followed by neutral lipid accumulation for biofuels.^{145, 146, 144, 148, 149 150} One study used a mixed culture composed of *Scenedesmus*, *Actinastrum*, *Chlorella*, *Spirogyra*, *Micractinium*, *Golenkinia*, *Chlorococcum*, *Closterium*, and *Nitzschia* for the treatment of dairy and wastewater in bench-scale reactors. While dairy wastewater resulted in lipid productivity of 17 mg/L/day, the municipal wastewater reactors resulted in 24 mg/L/day indicating that these mixed cultures were more efficient at assimilation of ammonia and phosphate into their biomass than production of lipids.¹⁴⁵

Lipid production

When considering doubling time and lipid accumulation, the two best strains for biodiesel production observed here were WC-1 (*Tetradesmus*) and WC-5 (*Chlorella*). The two highest lipid-producing strains were WC-5 (*Chlorella*) and UTEX395 (*Chlorella*). UTEX 395 was included as a control since it has previously been shown to be an oleaginous algae strain.^{151, 152} With fast doubling times, these two strains have a higher potential to outcompete contaminating, non-biodiesel-producing microorganisms. Algae strains with fast growth rates and high lipid content have the greatest potential to produce biodiesel.

WC-1 was previously identified as *Scenedesmus obliquus*¹¹⁵ and Gardner et al., (2012) showed WC-1 had a doubling time of 21 hours in an air only condition which is similar to the doubling time reported here at 19.81 ± 7.19 hours.¹⁵ The WC-1 DCWs ($0.99 \text{ g L}^{-1} \pm 0.07$; $\alpha=0.05$) were most similar to the DCWs reported by Gardner, 2012 ($1.1 \text{ g L}^{-1} \pm 0.1$; $\alpha=0.05$) which used the same 1.25 L photobioreactors, and aeration, allowing greater availability of CO₂ to the cells compared to algae grown in flasks.^{107, 115}

When screening for promising strains for lipid production, it can be important to determine which strains are responsive to sodium bicarbonate and which are not.^{69,92} Bicarbonate addition buffers the pH and provides higher concentrations of inorganic carbon necessary to make lipids. This combined with nitrate depletion, which reduces new biomass growth, can often lead to higher rates and extents of lipid production.^{49, 69, 70, 133} Not all algae increase lipid production with added sodium bicarbonate. In this study, there were four strains that did not respond to sodium bicarbonate addition. Gardner et al. (2012) found that when sodium bicarbonate was added to cultures *after* nitrate depletion, there was an overall loss in neutral lipids, whereas bicarbonate addition just before nitrate depletion increased lipid production significantly. Here, nitrate concentrations were measured and verified to be present in all strains prior to sodium bicarbonate addition. Strain WC-1 was previously reported to increase its neutral lipid content as a response to sodium bicarbonate addition.⁶⁹ Here the WC-1 medium nitrate concentration was present, but low ($4.3 \pm 2.5 \text{ mg L}^{-1}$). It is possible that WC-1 has a minimum medium nitrate concentration required for sodium bicarbonate addition to result in increased lipid content, or perhaps the nitrate was depleted completely between the time of nitrate measurement and the sodium bicarbonate addition. Further work is needed to elucidate when and why some strains increase lipid production with sodium bicarbonate addition and others do not.

It was hypothesized there would be a positive correlation between strains with low doubling times and high DCWs. When strains grow faster, they often accumulate more cells and ultimately more biomass. The results in Figure 3.9 indicate that there was no correlation observed between doubling times and DCWs for each strain as indicated by a linear regression. Similarly, it was hypothesized that there would be an inverse relationship for strains having low doubling times and high concentrations of lipids, as indicated by Nile Red fluorescence. When

strains grow more slowly, the fixed carbon may be stored as long-chain fatty acids rather than used to make new cells. However, the 11 algae strains characterized here showed an R^2 of 0.0207 indicating a lack of correlation between doubling time and Nile Red fluorescence (Figure 3.10).

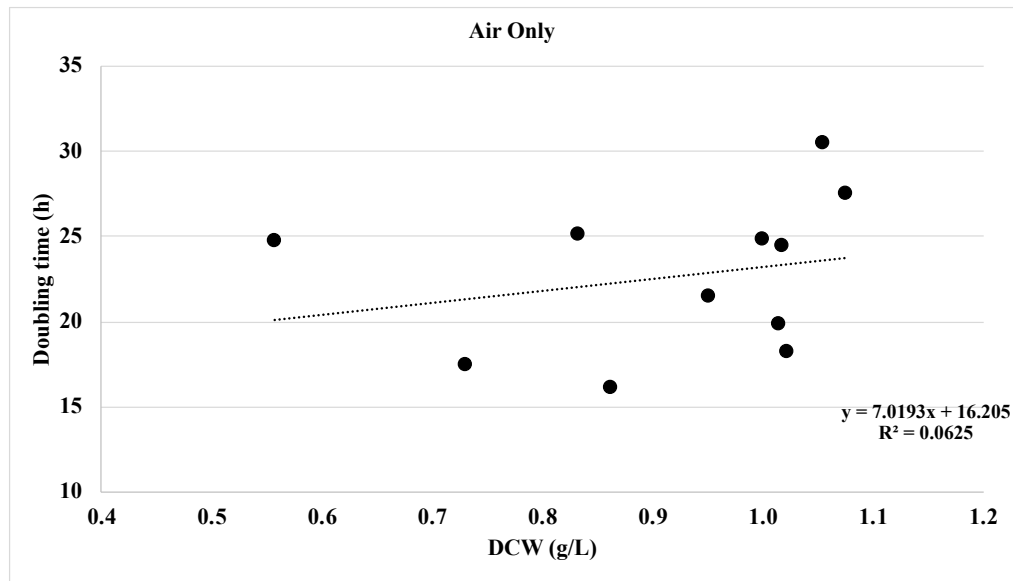


Figure 3.9 Linear regression of the doubling times and DCW for the 9 YNP green algae strains and 2 industrial strains. The R^2 for the linear regression was 0.0625 indicating that there was not a correlation between doubling time and DCW.

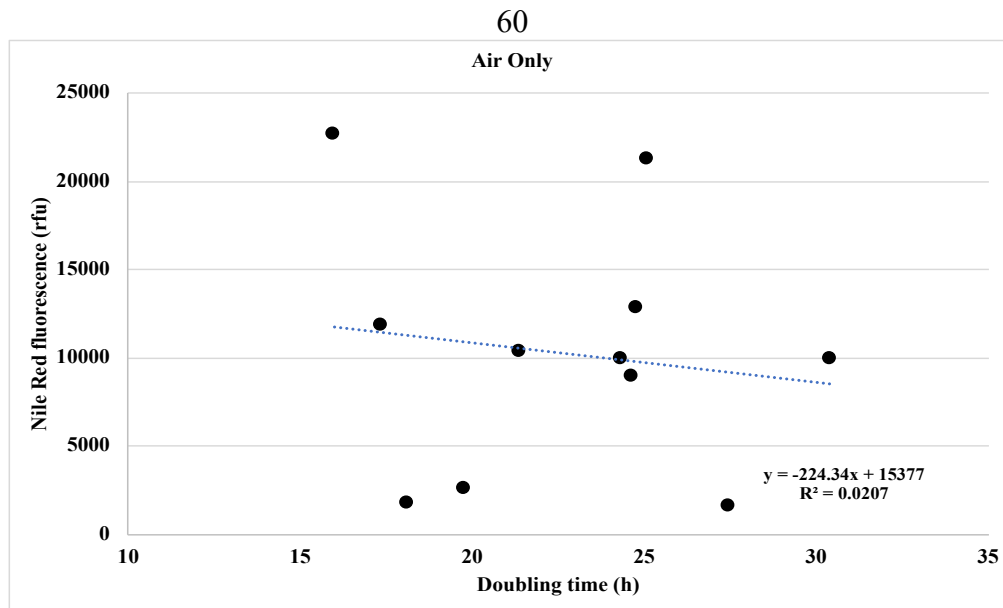


Figure 3.10 Linear regression of the doubling times and Nile Red fluorescence for the 11 YNP green algae strains. The R^2 for the linear regression was 0.0207 indicating that there was not a correlation between doubling time and Nile Red fluorescence.

The benefits of using extremophilic algae, especially alkaliphilic strains are numerous. In unbuffered systems, algae fix the acidic forms of dissolved inorganic carbon such as carbonic acid, driving the pH of the system to alkaline conditions.¹¹⁵ Algae that are adapted to alkaline conditions can withstand $\geq$ pH 8.5.¹⁰¹ Once the pH increases such that the DIC is limited to the alkaline CO_3^{2-} , it becomes harder for algae adapted to alkaline conditions to fix these species, so they stop fixing the DIC and the pH drops as CO_2 accumulates in the growth medium. At pH between 10.5 and 11, a pH-induced lipid accumulation “stress” can sometimes be observed.¹¹⁵ At pH > 8 , there is an increased flux of CO_2 into the growth medium increasing the inorganic carbon availability, which can be used for carbonaceous molecules such as TAGs, cellulose, and other carbon-rich molecules.^{49, 70, 107, 133, 144, 153}

Summary & Conclusions

In the work described here, 9 newly isolated extremophilic algal strains and 2 culture collection strains were compared and characterized for their potential for biomass and/or biofuel production. The results obtained represent initial detailed characterization, but not optimized growth conditions. Parameters such as temperature, pH, light intensity, and growth medium composition could further be optimized for each strain and will potentially improve growth rates, biomass yields, and the rate and extent of lipid accumulation.

Among the eleven strains tested, three were found to grow rapidly; WC-1, WC-5 and WC-2b. Two strains were identified as producing the highest DCW; PC-3 and WC-2b. Two strains were distinctly higher in lipid accumulation than the other nine; the two *Chlorella* strains, WC-5 and UTEX395. No single strain performed well in all three categories, however, two strains performed well in two categories; WC-5 performed well for growth and lipid accumulation and WC-2b had fast growth rates and high DCW density. Based on these results, of the 11 strains tested WC-5 and WC-2b are the best potential strains for further optimization of lipid for biodiesel and biomass production, respectively. Among the 9 green algae strains collected from one extremophilic stream, there are significant differences among them in terms of growth rate, biomass production and lipid accumulation.

62
CHAPTER FOUR

DRAFT GENOME FOR A NOVEL, EXTREMOPHILIC,
FRESHWATER DIATOM

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Karen M. Moll

Contributions: Performed data analysis, and writing.

Co-Author: Thiruvarangan Ramaraj

Contributions: Mentored data analysis

Co-Author: Nicholas P. Devitt

Contributions: Mentored data analysis

Co-Author: Connor Cameron

Contributions: Mentored data analysis

Co-Author: Matthew Fields

Contributions: Contributed to initial sequencing design and manuscript review

Co-Author: Joann Mudge

Contributions: Discussed results, mentored data analysis, manuscript review and revision.

Co-Author: Brent M. Peyton

Contributions: Discussed results, manuscript review and revision.

Manuscript Information

Karen M. Moll, Thiruvarangan Ramaraj, Nicholas P. Devitt, Connor Cameron, Matthew Fields,
Joann Mudge, Brent M. Peyton

BMC Genomics

Status of Manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☐ Published in a peer-reviewed journal

Abstract

Diatoms are known for atmospheric CO₂ remediation and decreasing ocean acidification. With changes in atmospheric CO₂, it is important to characterize diatom genomes as they are integral to understanding global carbon fixation, with potential for photosynthetic biodiesel production. Isolated from an alkaline stream (pH 9.3) in Yellowstone National Park (YNP), diatom strain RGd-1 has been shown in previous work to yield lipid concentrations up to 30-40% (w/w) triacylglycerol (TAG) and 70-80% (w/w) fatty acid methyl esters (FAMEs) that can be transesterified into biodiesel for biofuel applications. Here we report the 24 Mb draft genome for RGd-1, an extremophilic, freshwater, pennate diatom. RGd-1 was found to align best to the centric diatom, *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* on a nucleotide and protein level, respectively. A *de novo* transcriptome assembly was used to annotate the RGd-1 genome assembly. RGd-1 was shown to have a nearly complete glyoxylate pathway that could be used as a carbon conservation strategy to accumulate high concentrations of neutral lipids. As part of the RGd-1 whole-genome sequencing project, we assembled an associated, novel 3.1 Mb *Brevundimonas* sp. genome. Nine major bacterial OTUs were found in the RGd-1 culture through 16S amplification and sequencing. Of those strains, seven may produce iron chelating siderophores, which could make iron biologically available to RGd-1 in an alkaline environment.

Diatoms, unicellular, photosynthetic algae with siliceous cell walls, are critical to ecosystem health at a global scale and have the capacity to help buffer climate change. Producing at least 25% of Earth's atmospheric oxygen, diatoms also fix ~25–45% of global CO₂ directly mitigating the major driver behind climate change.¹² However, diatoms are not immune to the effects of climate change, manifested in aquatic systems as rapid temperature fluctuations and acidification. As a result, a drastic global decline in diatoms has recently been observed that may directly reduce oxygen contributions and carbon fixation, further compounding climate change.¹²

As aquatic phytoplankton (class Bacillariophyceae) that contribute to primary production, diatoms fix dissolved CO₂ as a necessary requirement for growth. The innate ability of diatoms to sequester dissolved CO₂ and incorporate the carbon into either biomass, starch, or lipids, makes these unicellular microorganisms suitable candidates for large-scale cultivation in terms of CO₂ utilization and valuable by-products. Once diatoms die, their siliceous cell walls sink to the bottom of the ocean, thus removing CO₂ from the atmosphere and adding a carbon source for the utilization of benthic organisms.^{154, 155} They are especially promising given their potential to outcompete other phytoplankton based on silicate availability.¹⁵⁴ Accordingly, diatoms may be viewed as natural sources of carbon sequestration.¹⁵⁴

While diatoms have been largely overlooked in biofuel research, which has primarily focused on green algae, there are many good reasons to consider diatoms for biofuel research and potential production. As a near carbon-neutral technology, using phototrophs for biofuel production would limit additional CO₂ emissions while helping to meet transportation fuel demands. Further, in addition to high lipid accumulation, diatoms can amass high concentrations of other carbonaceous compounds useful for the production of renewable fuels and high-value

co-products (e.g. chrysolaminarin, and fucoxanthin).^{15, 156} With these attributes, diatoms can help recycle atmospheric carbon dioxide while producing biofuels, thereby contributing to improved environmental health. Understanding the genetic potential of these understudied organisms is now more critical than ever. Characterization of the genome sequence has revealed important information on the metabolic capacity that provides an important foundation for monitoring or manipulating biofuel production capabilities.

RGd-1, a pennate diatom most closely related to *Lemnicola hungarica* based on SSU rDNA (18S), was isolated from Witch Creek, an alkaline (pH 9.3), thermally-impacted, arsenic-containing (300 ppb) creek in Yellowstone National Park (YNP).⁴⁹ Previously, RGd-1 was shown to accumulate very high concentrations of triacylglycerol (TAG) 30–40% wt/wt and fatty acid methyl ester (FAME) 70–80% wt/wt, making it a promising candidate for biofuel production.⁴⁹

Until recently, most diatom biofuel research has been performed on axenic strains.^{10, 11, 157, 158} Recent research has shown there are benefits for established relationships between bacteria and diatoms.^{44, 159, 160} While some bacteria associated with diatoms have been identified, specific interactions are much less well defined⁴⁴ with the literature being particularly scarce for microbiomes of non-marine diatoms. Bacteria attached to diatom frustules have been found to exchange substances such as vitamins, indole, and organic carbon compounds with their associated diatom and to receive protection from predators.¹⁶¹⁻¹⁶⁴ Bacteria attached to diatoms have been hypothesized to promote organic matter degradation thus increasing community biomass due to cross-feeding and increasing sedimentation, which ultimately provides organic carbon to benthic communities.^{165, 166} Additionally, while diatoms do not produce siderophores (iron-chelating molecules that bind iron in bioavailable forms) it is thought that siderophore-

producing bacteria may provide bioavailable iron to diatoms, a mechanism that is particularly important in alkaline environments where bioavailable iron is scarce.^{44, 159}

The association of bacteria with diatoms can be thought of as a “diatom microbiome”. The bacteria reside in the “phycosphere,” defined by Bell and Mitchell as a region around an algal cell surrounded by bacteria.⁴² This is an important concept in the study of algae and is becoming increasingly appreciated. Lending support to the importance of the diatom phycosphere, several marine diatoms have been observed to harbor distinct bacterial communities.^{44, 161, 167} For instance, bacteria attached to *Thalassiosira rotula* and *Skeletonema costatum* have been found to consistently be members of the *Flavobacter-Sphingobacteria* groups within the Bacterioidetes phylum and the unattached bacteria are commonly in the *Roseobacter* group within α -Proteobacteria.¹⁶¹ As an extremophilic, freshwater diatom, RGd-1 was found to contain bacteria remarkably different from those that have been associated with marine diatoms.

Given the unique environment of RGd-1 and the lack of available diatom genomes, the RGd-1 diatom was selected for whole-genome sequencing to obtain insight into unique characteristics and its ability to accumulate high concentrations of lipids.⁴⁹ Further, it was possible to gain insight into how a fresh-water diatom adapted to extreme conditions of temperature, pH, and metals. Here, we present the draft genome assembly for the extremophilic, freshwater diatom, RGd-1, and associated bacteria.

DNA Extraction

RGd-1 high molecular weight DNA was extracted using a cetyl trimethyl ammonium bromide (CTAB) DNA extraction method (see Appendix B)¹¹⁶ and amplified according to the JGI DNA extraction protocol.¹⁶⁸ The RGd-1 culture was unialgal, but even after antibiotic treatments with Ampicillin (1000 mg L⁻¹), a beta-lactam antibiotic that inhibits cell wall synthesis,^{169, 170} and Imipenem (5 mg L⁻¹), a fluoroquinolone that inhibits prokaryotic DNA gyrase (topoisomerase),¹⁷¹ the culture still contained bacteria.

Whole-genome Sequencing

Illumina

Illumina® (San Diego, CA) HiSeq 2000 2x50 reads were sequenced at the National Center for Genome Resources (Santa Fe, NM). TruSeq DNA libraries were prepared using a 300 bp insert size. A total of 84,132,116 paired-end, 2x50 reads were generated for a total of 366x coverage.

PacBio

Long-reads were obtained from the Pacific Biosciences (Menlo Park, CA) sequencing technology at the National Center for Genome Resources. A total of 11 single-molecule real-time Single Molecule Real Time (SMRT) cells were used for sequencing. Seven SMRT cells were run on the Real-time Sequencing (RS)I and four SMRT cells were run on the RSII machine. The first nine SMRT cells used chemistry 2 polymerase 4 (C2P4) (v. 1.0) and the remaining three were run with C4P6 chemistry (6-hour movies) to improve read length and coverage (v. 1.5). BluePippen

size selection at 10Kb lengths was used for sample preparation prior to sequencing.¹⁷² The PacBio sequencing produced a total of 226,481,107 raw reads and 206,073,836 filtered reads yielding 1,723,858,582 bp for a total of 74X coverage.

Assembly Methods

The PacBio reads were assembled using the de Bruijn graph to overlap consensus (DBG2OLC) read assembler,¹⁷³ which is a hybrid assembler that uses Illumina reads to error-correct longer PacBio reads, which are then assembled. The PacBio reads were also assembled separately, using Falcon¹⁷⁴ (RGd-1 v. 1.0) or Canu (RGd-1 v. 1.5).¹⁷⁵ The DBG2OLC assembly was then used as “trusted contigs” in SPAdes with the Illumina reads to error-correct. The SPAdes assembly was then filtered for contigs ≥ 1000 bp,¹⁷⁶ after which CAP3 was used to improve scaffolding¹⁷⁷ and as a clustering algorithm; and CD-HIT-EST was used to remove redundancies.¹²⁵ Finally, Gapcloser (part of the SOAPdenovo package) was used to fill gaps present in the scaffolds (v. 1.0. v. 1.5).¹⁷⁸ An additional assembly incorporated a small percentage of long PacBio reads that were used for scaffolding (v. 1.5). The alternative assembly, v. 1.5, v. 1.0 was used as “trusted contigs” in SPAdes and all other downstream steps remained the same for the two assemblies.

BUSCO

Benchmarking Universal Single Copy Orthologs (BUSCO) software provides a quantitative assessment of genome assemblies based on orthologs selected from OrthoDB v.9, a catalog of orthologous protein-coding genes for vertebrates, arthropods, fungi, plants, and bacteria.¹⁷⁹ Assembly assessments were performed using the eukaryote lineage within BUSCO 3, which contains 303 genes that are present in at least 90% of the eukaryotic species used to

assemble the database, though other lineages were also tried.³¹ BUSCOs were identified by tBLASTn searches that align proteins against translated DNA using the basic local alignment search tool (BLAST) followed by Augustus gene predictions^{180, 181}, and classified into lineage-specific matches using the HMMER program within the BUSCO package.

Read Alignments and Validation

Illumina 2x50 reads were aligned to the genome assembly using the Burrows-Wheeler Aligner (BWA) algorithm (version 0.7.12) with a maximum of two paths and sequence alignment map (sam) output format to determine how well the genome was assembled.¹⁸²

Structural Annotation

The MAKER genome annotation pipeline was used for structural annotation of the RGd-1 genome.¹⁸³⁻¹⁸⁷ The MAKER pipeline aligns the provided ESTs to the genome and creates *ab initio* gene predictions with SNAP¹⁸⁸ and Augustus¹⁸⁰ using evidence-based quality values. Following the completion of MAKER, the files were combined using `fasta_merge` and `gff3_merge` that are included as part of the MAKER package to obtain cumulative fasta and gff files that represent the annotated genes in the genome assembly.

Functional Annotation

The transcriptome assembly was used as EST-evidence within Maker.^{186, 187} The transcriptome was filtered for reads ≥ 500 bp. The resulting proteins were used as input for the web-based, functional annotation program, GENAS (v. 6.0).¹⁸⁹ The MAKER predicted proteins were used as input into KeggMapper for annotation of metabolic pathways.¹⁹⁰

K-mer Analysis

A k-mer sweep was performed on the Illumina 2x50 nt short-reads using KAT.^{33, 191} KAT is a reference-free toolkit that uses k-mer frequencies and GC content to assess the quality of genome assemblies.³⁴

Concatenated Protein Phylogenetic Tree

A concatenated protein tree was created to more accurately determine the phylogeny of RGd-1 through modification of the ezTree pipeline.¹⁹² Protein sequences of 18 publicly available algal genomes were downloaded from the JGI (<https://genome.jgi.doe.gov>) and ensemble genome portals (<http://ensemblgenomes.org>). The ezTree pipeline was modified to skip the gene prediction step, and the single-copy marker proteins that were shared among all genomes were identified with the reference database, Pfam v32.0¹⁹³ (Supplemental table B.3). These proteins were aligned individually with MAFFT-L-INS-I,¹⁹⁴ and trimmed by trimAL using a 75% gap threshold.¹⁹⁵ The final proteins were concatenated to generate the final alignment file and viewed with FigTree.¹⁹⁶

16S Amplified Sequencing and Analysis

For the associated bacteria, 16S amplified sequencing was performed post-antibiotic treatments. The SSU rRNA gene was amplified using FD1 (5'-AGAGTTTGATCCTGGCTCAG-3') and 529R (5'-CGCGGCTGCTGGCAC-3'), targeting the V1-V3 region of bacteria.^{122, 197} High-quality amplicons were sequenced on the Illumina[®] MiSeq platform (Illumina[®], San Diego, CA) according to the manufacturer's procedure for paired-end sequencing with 300 cycles.¹⁹⁸ Two PCR reactions were performed. The first PCR reaction was performed to amplify the region of interest after which amplicons were purified with Ampure beads. A second PCR

was performed to index each sample to ensure sample identification for analyses. Indexed samples were again purified with Ampure beads and target concentrations were determined with Pico-green 480 nm/520 nm excitation/emission filters (Quant-IT, Invitrogen, Carlsbad, CA, USA) and measured on a fluorescent plate reader (BioTek, Synergy H1 Hybrid Reader). Final sample concentrations were normalized so that the same concentration of each sample was added. The PhiX control library was added at 10% of the sample DNA concentration. The samples and control library were pooled and sequenced.¹⁹⁸ For analysis, reads were analyzed using Clark, a k-mer based clustering algorithm¹⁹⁹ using the ribosomal database project (RDP) 16S SSU database for identification.²⁰⁰

RNA Sequencing and Transcriptome Assembly

RNA from 9 discreet RGd-1 samples from three different experimental conditions, each run in biological triplicates, were extracted and purified using the NEB Monarch Total RNA Miniprep Kit, according to the manufacturer's instructions. The samples of extracted mRNA were submitted for strand-specific RNA sequencing. Illumina® (San Diego, CA) HiSeq 5000 2x150, single-index reads were sequenced at Genewiz (South Plainfield, NJ). TruSeq Stranded libraries were prepared using a 300-400 bp insert size. PolyA selection was used for the removal of rRNA. A total of 424,066,367 paired-end reads and 127,218 Mb yield with a mean quality score of 38.05 (provided by the Genewiz NGS Data Report).

Transcript Alignment and Assembly

Transcripts were aligned with the strand-aware aligner, HiSat2, to determine how well the reads aligned against the RGd-1 assembly.²⁰¹ The reads were trimmed using the Trimmomatic function within the Trinity assembler after which the reads were assembled in *de-*

novo mode and filtered for lengths ≥ 500 bp.^{184, 185, 202} To determine the best transcriptome assembly method, the trimmed reads were also assembled against RGd-1 v. 1.0 in a reference-guided transcriptome assembly in genome-guided mode within Trinity.²⁰²

Results

The RGd-1 genome assembly was created using 366X Coverage of Illumina 2x50 reads and 74X coverage of PacBio reads. As shown in Table 4.1, the RGd-1 genome assembly consisted of 520 contigs with a total length of 23.3 Mb. The longest contig was 372,285 bp with a contig N50 of 102,387 bp. Using the genome annotation tool, MAKER, the RGd-1 genome assembly was found to have 13,422 gene models.

Table 4.1 Genome assembly statistics for two RGd-1 assembly versions, v. 1.0 and v. 1.5 (with a small percentage of additional long PacBio reads).

Statistic	RGd-1 (v 1.0)	RGd-1 (v 1.5)
Number of contigs	520	1537
Genome length	23,268,988	27,896,030
Contig N50	102,387	37,952
Maximum contig size	372,285	243,262
Minimum contig size	1,041	1,294

Assembly Comparison

Two different RGd-1 draft genome assemblies were compared to determine the best quality assembly (v. 1.0) to proceed with further analysis (Table 4.1). The primary difference between the two assemblies was the incorporation of additional C4P6 PacBio reads. However, the majority of the PacBio reads from the C4P6 PacBio sequencing project were bacterial, and only approximately 5% of the total reads were assigned to class Bacillariophyceae to which

diatoms belong. The diatom reads were used for scaffolding. However, all attempts to include the small fraction of long, diatom PacBio reads resulted in three times the number of contigs and approximately 1/3 the contig N50 (v. 1.5) compared to the original assembly v. 1.0 (Table 1.0) despite there being a larger number of long contigs (Figure 4.1).

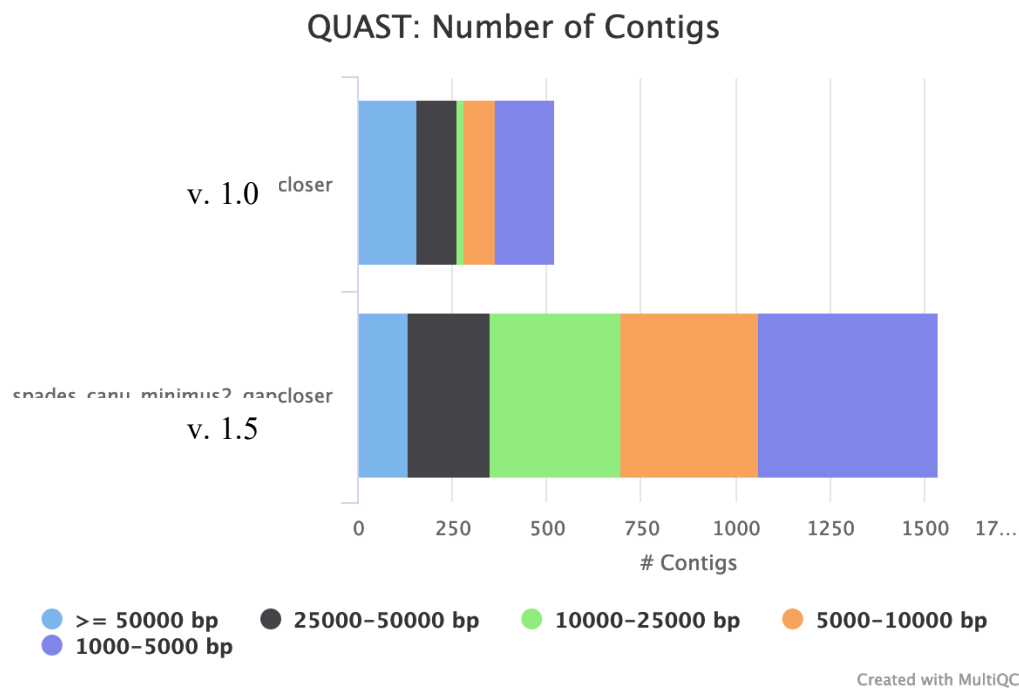


Figure 4.1 Comparison of two draft RGd-1 genome assemblies, v. 1.0 and v. 1.5. The difference between the two assemblies was the inclusion of an additional small PacBio dataset. This figure was generated using MultiQC.²⁰³

Gene Space Completeness

To assess genome assembly completeness, or how well the genome was constructed in terms of genes captured, the assembly was evaluated with BUSCO. BUSCO assumes the orthologous groups it uses will be identified in 90% of the organisms within that lineage. Given the primary and secondary endosymbiotic events that have occurred within the diatom group,²⁰⁴ many genes in either the protist or eukaryota lineage may have diverged enough that they were

not identified as complete, partial, or fragmented genes, but were instead grouped by BUSCO into the “missing genes” category. The RGd-1 assembly was analyzed using the eukaryota, chlorophyta, alveolata, and protist lineages (Table 4.2). The best results were generated when BUSCO was run with the eukaryota lineage, which was then used for further comparison. The eukaryote lineage found 57.8% single copy, 1.0% duplicate, 6.6% fragmented, and 34.6% missing orthologs to be captured in the genome (Table 4.3). The large percentage of missing orthologs may be due to the incomplete genome assembly but may also reflect evolutionary distance from the gene models included in the eukaryota lineage. In other words, some of the missing orthologs may be present but have evolved enough to not be recognized.

Table 4.2 The RGd-1 v.1.0 genome assembly was analyzed using five BUSCO lineages, Eukaryota, Protists, Alveolata/Stramenopiles, Chlorophyta, and Embryophyta. The gene capture percentage was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies.

BUSCO type	Eukaryota	Protists	Alveolata Stramenopiles	Chlorophyta	Embryophyta
	odb9	ensemble	ensemble	odb10	odb9
Complete	58.8%	46.1%	26.1%	21.9%	7.5%
Complete-single copy	57.8%	45.6%	26.1%	21.0%	6.9%
Complete-duplicated	1.0%	0.5%	0.0%	0.9%	0.6%
Fragmented	6.6%	0.5%	0.4%	3.5%	0.8%
Missing	34.6%	53.4%	73.5%	74.6%	91.7%
Total orthologs searched	303	215	234	2168	1440

Table 4.3 Gene capture measured by BUSCO. A total of 303 BUSCOs were searched within the eukaryota lineage.

BUSCO type	Number of BUSCOs found	% BUSCOs found
Complete-single copy	175	57.8%
Complete-duplicated	3	1.0%
Fragmented	20	6.6%
Missing	105	34%

To distinguish between these two possibilities for missing orthologs due to an incomplete genome assembly or evolutionary distance, the *P. tricornutum* genome assembly was analyzed with the eukaryota lineage. *P. tricornutum* has a quality genome assembly where the number of scaffolds (33) is equivalent to the number of chromosomes³⁵ and most genes are expected to be captured. We would, therefore, expect orthologues labeled as missing to be present but evolved beyond recognition. *P. tricornutum* generated 75.0% complete orthologs, 6.6% fragmented, and 20.4% missing (Table 4.4). While the number of complete orthologs is 17% higher for *P. tricornutum* than RGd-1, indicating that there is room for assembly improvement, there are still 20% missing orthologs in the *P. tricornutum* genome assembly, indicating that the eukaryota BUSCO lineage does not capture unique features within the diatom lineage.

To determine how well BUSCO captures diatom genes across different genomes, four additional publicly available diatom genomes were also evaluated. Like *P. tricornutum*, *T. pseudonana* had a high percentage of captured complete orthologs (65.7%). *C. cryptica*, *P. multiseriis*, and *F. cylindrus* had successively lower complete orthologs at 61.1%, 59.4%, and 40.6%, respectively (Table 4.4). The minimum percentage of missing orthologs was 20.4% for *P. tricornutum*, which represents the “best-case scenario” of the available diatom genome assemblies since it is regularly updated in ensemble. To reiterate, the comparison of BUSCO results across six genome assemblies indicates that at least 20% of the 303 genes present in 90%

of the species within the eukaryota lineage are divergent from the diatom lineage. If we expect to capture at best 75.0% of the diatom genes, the recalculated BUSCO scores for RGd-1 are 72.6% single copy, 1.2% duplicate, 8.6% fragmented orthologs within the eukaryota lineage.

Table 4.4 RGd-1 v.1.0 genome assemblies were compared to the other publicly available diatom genome assemblies, *P. tricornutum*, *T. pseudonana*, *C. cryptica*, *P. multiseriis* and, *F. cylindrus*. The gene capture percent was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies. A total of 303 BUSCOs were searched within the eukaryota lineage.

BUSCO type	RGd-1 v.1.0	<i>P.</i> <i>tricornutum</i>	<i>T.</i> <i>pseudonana</i>	<i>C.</i> <i>cryptica</i>	<i>P.</i> <i>multiseriis</i>	<i>F.</i> <i>cylindrus</i>
Complete- single copy	57.8%	73.3%	65.7%	61.1%	59.4%	40.6%
Complete- duplicated	1.0%	1.7%	1.7%	5.6%	1.0%	65.7%
Fragmented	6.6%	4.6%	7.9%	7.3%	6.3%	61.1%
Missing	34.6%	20.4%	24.0%	26.0%	33.3%	59.4%

Transcriptome Assembly

The raw transcripts aligned to the RGd-1 genome assembly with an alignment rate of 20.96%. When the reads were trimmed, the percent alignment increased to 33.31% (Figure 4.2). This alignment is low and potentially indicates the presence of bacterial contaminants, despite the RNA depletion step completed as part of the Illumina library preparation.^{205, 206} Nevertheless, the *de-novo* transcript assembly captured 83.8% complete (27.7% single-copy, 56.1% duplicated) with 5.0% fragmented and 11.2% missing orthologs. A total of 11,150 proteins were identified following the GENSAS EvidenceModeler annotation function.¹⁸⁹

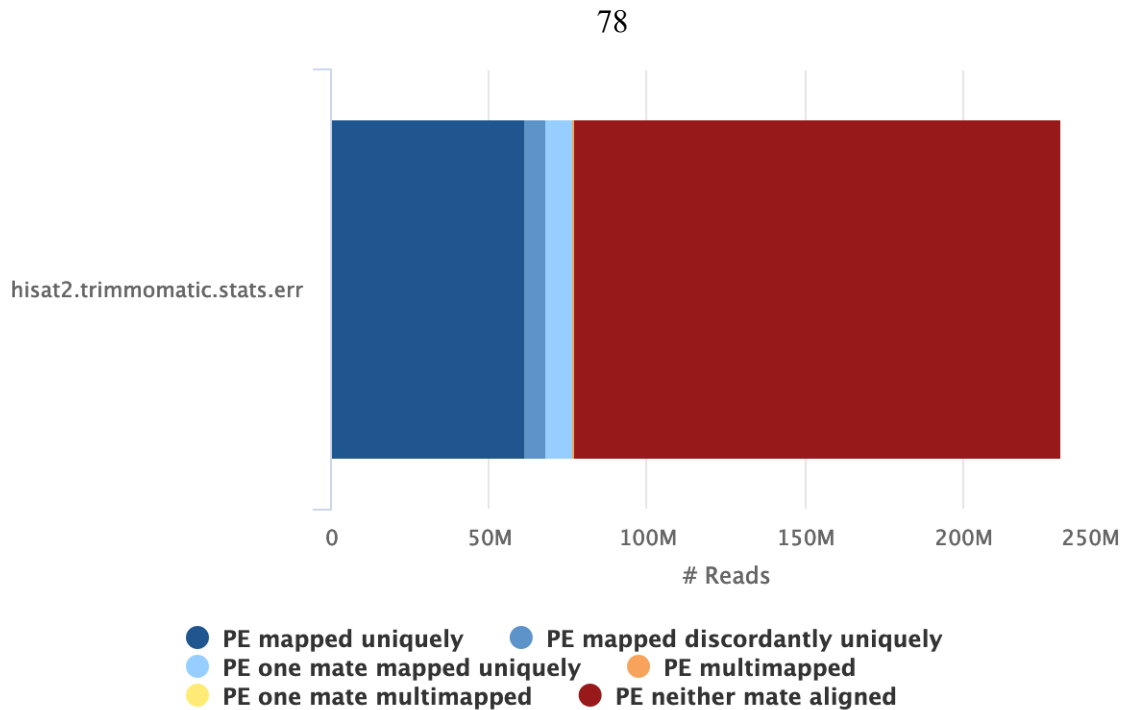


Figure 4.2 The number trimmed paired-end reads that were uniquely aligned as pairs, had one mate pair uniquely, one mate mapped in multiple locations, the pairs mapped discordantly, the pair-end reads mapped in multiple locations, or neither read aligned. Reads were aligned using HiSat2 and the figure was generated using MultiQC.²⁰³

To confirm the best method to assemble the RGd-1 transcriptome, the *de novo* assembly was compared to a reference-guided assembly using the RGd-1 v. 1.0 genome assembly. The transcriptome assembly statistics in Table 4.5 show pre- and post-filtering for reads ≥ 500 bp. While the filtered *de novo* assembly had an 82% increase in the number of contigs, compared to the filtered reference-guided assembly, the *de novo* genome length was nearly 50% longer, indicating that a greater number of genes could be captured (Tables 4.5, 4.6).

Table 4.5 Genome assembly statistics for two RGd-1 transcriptome assembly versions, *de novo* and reference-guided using the RGd-1 v.1.0 genome assembly. Each transcriptome assembly shows the statistics pre- and post-filtering for reads ≥ 500 bp.

Statistic	<i>de novo</i>	<i>de novo</i> filtered	Reference-guided	Reference-guided filtered
Number of contigs	1,875,787	78,720	222,478	42,868
Genome length	496,220,882	142,135,629	124,808,503	74,549,208
Contig N50	225	3,052	845	2,853
Maximum contig size	29,652	29,652	24,608	24,608
Minimum contig size	129	500	180	500

A comparison between the BUSCO results for the *de novo* transcriptome assembly and reference-guided assembly found 11.2% more complete BUSCOs for the *de novo* assembly. Furthermore, the reference-guided assembly had 10.3% more missing orthologs than the *de novo* assembly. While there are more complete-duplicated orthologs in the *de novo* assembly, they are still within the complete fraction and may represent true duplications.²⁸ Based on the BUSCO results combined with the assembly statistics, the *de novo* transcriptome assembly was used for annotation.

Table 4.6 Two different RGd-1 transcriptome assemblies were compared; a *de novo* assembly and reference-guided assembly. The gene capture percent was measured as a fraction of the total number of searchable BUSCOs identified in the assemblies. A total of 303 BUSCOs were searched within the eukaryota lineage.

BUSCO type	<i>de novo</i> assembly	Reference-guided assembly
Complete total	83.8%	72.6%
Complete-single copy	27.7%	35.3%
Complete-duplicated	56.1%	37.3%
Fragmented	5.0%	5.9%
Missing	11.2%	21.5%

Assembly Completeness

To validate the RGd-1 genome assembly, raw Illumina 2x50 nt reads were aligned against the genome assembly using the Burroughs Wheeler Aligner (BWA)²⁰⁷ algorithm to estimate the percentage of the genome that was assembled. The Illumina capture percent was 74.9% across the entire assembly, indicating that approximately 25% of the reads did not align to the genome, likely reflecting missing parts of the assembly, large repetitive regions that are not expected to be captured by the short reads or contaminating bacterial sequences.

A K-mer sweep can reveal insight about a genome assembly including contamination, and genome complexity.^{34, 208} As shown in Figure 4.3, the sweep revealed two peaks with the coverage (the number of reads containing a specific kmer, here kmer-27 was used for the kmer sweep) of the rightmost peak approximately two times the size that of the left peak. The left peak likely represents sequences from heterozygous regions in which haplotypes generated different k-mers with half as much coverage as those in the right peak, which were likely derived from homozygous regions (Figure 4.3).

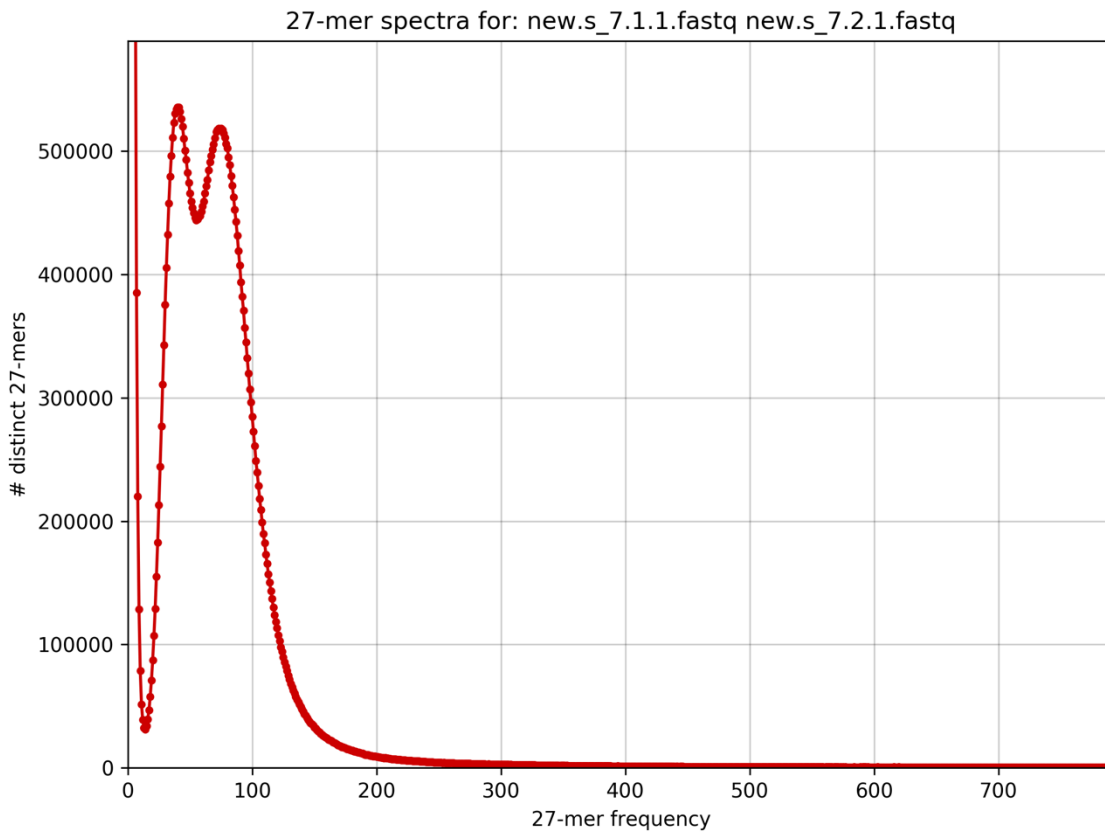


Figure 4.3 A k-mer sweep generated by KAT³⁴ using a k-mer length of 27. The analysis was performed using the paired-end, 50 bp, Illumina reads.

Comparative Genomics

The diatoms, *C. cryptica*, *F. cylindrus*, *P. multiseriis*, *P. tricornutum*, and *T. pseudonana* were each aligned to the RGd-1 genome assembly to determine the percentage of nucleotides in common across their entire genomes (Table 4.7).

Table 4.7 Whole-genome alignments using BWA mem where the RGd-1 v. 1.0 genome assembly was indexed and the other assembly was queried. The publicly available diatom genomes, *C. cryptica*, *F. cylindrus*, *P. multiseriis*, *P. tricornutum*, and *T. pseudonana* were each aligned to RGd-1 on the nucleotide-level.

Genomes Aligned		% Aligned
RGd-1 v. 1.0	<i>C. cryptica</i>	3.29
RGd-1 v. 1.0	<i>F. cylindrus</i>	13.38
RGd-1 v. 1.0	<i>P. multiseriis</i>	0.59
RGd-1 v. 1.0	<i>P. tricornutum</i>	5.0
RGd-1 v. 1.0	<i>T. pseudonana</i>	33.33

There were generally low overall alignment percentages on the nucleotide level between RGd-1 and the marine diatoms, with the highest being *T. pseudonana* at 33.33% on a whole-genome level. This is surprising given that *T. pseudonana* is a centric diatom and RGd-1 is a pennate diatom. The centric and pennate diatoms are thought to have diverged approximately 130 million years ago during the Cretaceous Period.²⁰⁹ Given the extremophilic, freshwater habitat of RGd-1, it is possible that RGd-1 and potentially other extremophilic freshwater diatoms have greater divergence from the marine pennate lineage than previously recognized.

Using a concatenated protein tree to determine phylogeny consisting of 15 publicly available algae including; 1 red alga, 1 brown alga, 10 green algae and 3 diatom genomes and RGd-1, RGd-1 was found to be most similar to *P. tricornutum* and *P. multiseriis* (Figure 4.4) which is compatible with the shared pennate morphology of RGd-1, *P. tricornutum*, and *P. multiseriis* which may be a reflection of different adaptations that have occurred as a result of the different evolutionary pressures from their respective environments. This result differs from the whole genome alignment on a nucleotide-level, which indicated the most shared nucleotide space with *T. pseudonana*.

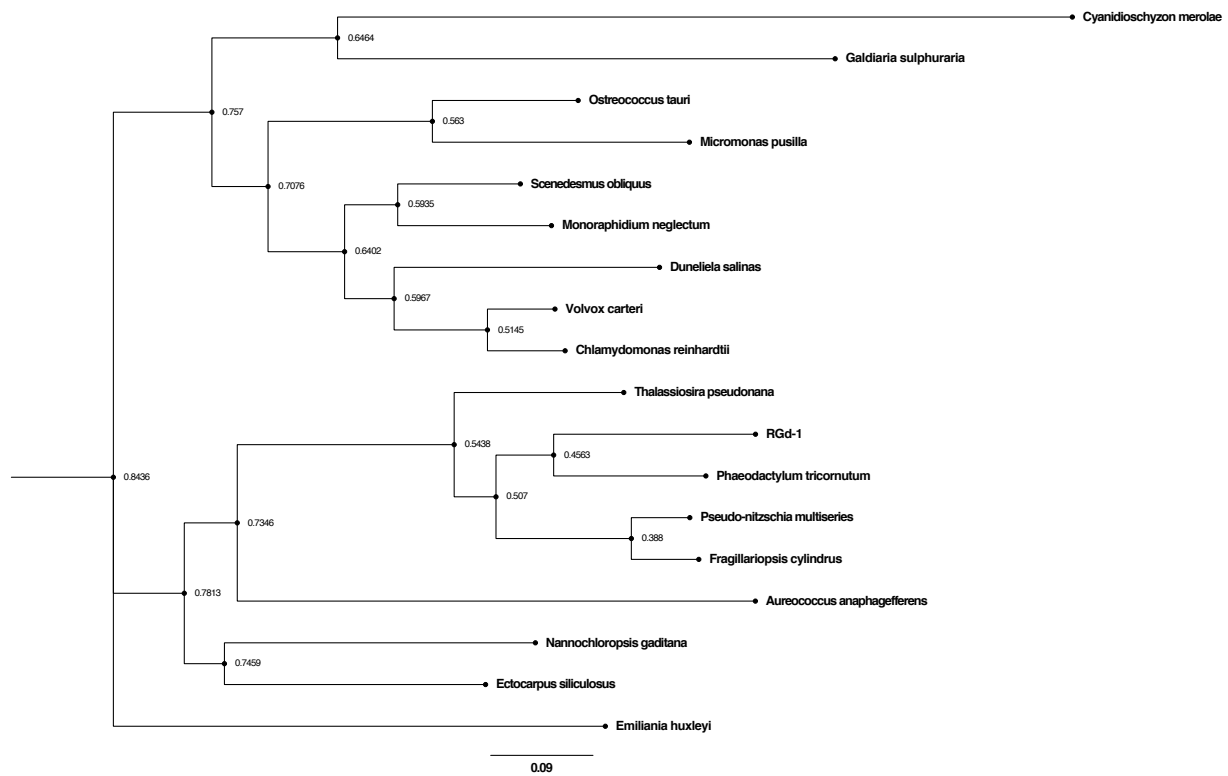


Figure 4.4 The protein sequences from fifteen publicly available genome annotations (1 red alga, 1 brown alga, 11 green algae and 3 diatoms) and RGd-1 were used to construct a concatenated protein using a modified ezTree pipeline.¹⁹² RGd-1 was phylogenetically closest to *P. tricornutum* on a protein-level. The proteins were trimmed using trimAL¹⁹⁵ and aligned with MAFFT-L-INS-i.¹⁹⁴ The scale represents the bootstrap values.

Figure 4.5 shows the extent to which RGd-1 and *P. tricornutum* are related on a nucleotide and protein level. The x-axes contain the *P. tricornutum* assembly chromosomes. The y-axes contain the 520 current contigs in the RGd-1 assembly. If *P. tricornutum* and RGd-1 were perfectly syntenous, the slope of the line would be 1.0. As shown in Figure 4.5, RGd-1 differs significantly on a nucleotide level, but is more syntenic on a protein level. The lack of synteny between *P. tricornutum* and RGd-1 indicates that on a nucleic acid level, the two diatoms have evolutionarily diverged. However, on a protein level, the two diatoms have more conservation, indicating that shared proteins may have similar functions.

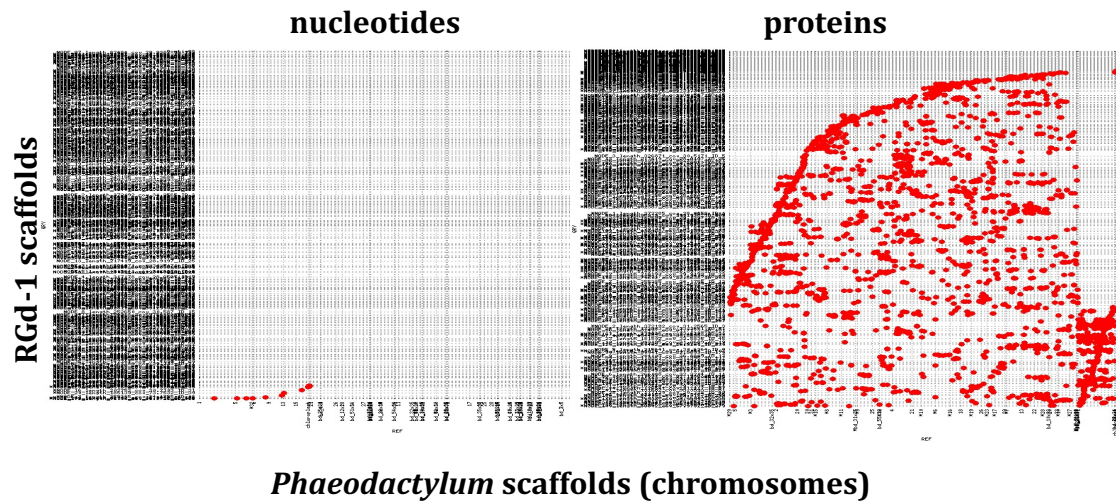


Figure 4.5 Comparison of *P. tricornutum* and RGd-1 assemblies based on amino acid sequences. The x-axis contains the scaffolds for the *P. tricornutum* assembly and the y-axis contains the 520 scaffolds for the RGd-1 genome assembly. Within the mummer packager, promer translates the nucleic acid-based assemblies into amino acids.²¹⁰ Perfectly syntenous assemblies would have a slope of 1.0.

RGd-1 Genome-based Metabolic Pathway Analysis

Annotated pathways were approximately 80% complete, as determined by the number of genes found within each pathway. This matches the BUSCO scores with complete total orthologs of 83.8% for the *de novo* transcriptome assembly that was used for annotation. The missing genes may be explained by an incomplete assembly or because the genes are evolutionarily distant from currently known genes in the databases.³¹

RGd-1 was found to have a nearly complete carbon fixation pathway (Figure 4.6). There is evidence that RGd-1 uses C₄-metabolism and C₃-metabolism in Figure 4.6. C₄ metabolism is more energetically expensive than C₃ metabolism. However, C₄ metabolism is more photosynthetically efficient.⁶⁰ Organisms that use C₄-metabolism use the 4-carbon intermediate, oxaloacetate as the primary product of carbon fixation. C₄ metabolism differs from C₃

metabolism by the separation of two intracellular compartments, allowing for two carboxylations. There is evidence that the primary carboxylation for the *T. pseudonana*, *P. tricornutum* and *F. cylindrus* occurs by the carboxylation of HCO_3^- with phosphoenolpyruvate carboxylase (PEPCase) resulting in oxaloacetate and inorganic phosphate in the outer compartment of the chloroplast.²⁰⁴ Oxaloacetate is transformed into another four-carbon acid such as malate or aspartate and transported from the outer compartment to the inner compartment of the chloroplast.²⁰⁴ The second carboxylation occurs when the CO_2 is fixed by Rubisco and the Calvin-Benson cycle. Following decarboxylation of the four-carbon intermediate, a three-carbon intermediate such as pyruvate or alanine are released to the outer compartment of the mitochondria. The results presented here indicate that pyruvate-phosphate dikinase (PPDK) to produce phosphoenolpyruvate with aspartate as an intermediate. Aspartate moves back into the mitochondria where it is converted to oxaloacetate by aspartate aminotransferase.²⁰⁴ These results are consistent with the analysis by Valenzuela et. al. 2012 on *P. tricornutum*.⁶⁷

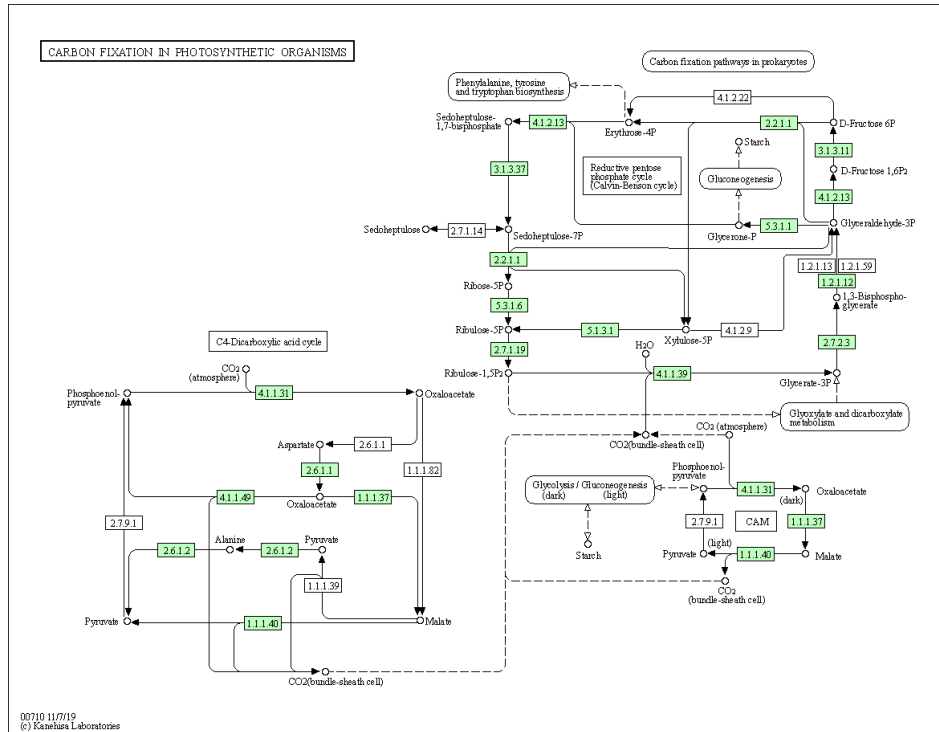


Figure 4.6 Annotated carbon fixation pathway in photosynthetic organisms. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

Figure 4.7 shows many core RGd-1 genes present in the TCA cycle. However, the following genes were not identified; aconitase and isocitrate dehydrogenase. Interestingly, isocitrate lyase was identified in the glyoxylate and dicarboxylate pathway (Figure 4.8) where isocitrate is converted to succinate with glyoxylate as an intermediate, followed by the presence of malate synthase to form malate. The carboxylation of malate then forms malonyl-CoA via acetyl-CoA-carboxylase after which the malonyl-CoA feeds into fatty acid biosynthesis. Overall, the glyoxylate and dicarboxylate pathway bypasses two decarboxylations that normally occur in the TCA.²¹¹ It may be possible that the use of the carbon conserving glyoxylate shunt that bypasses removal of CO₂, may partly explain why RGd-1 can accumulate high concentrations of lipids.⁴⁹

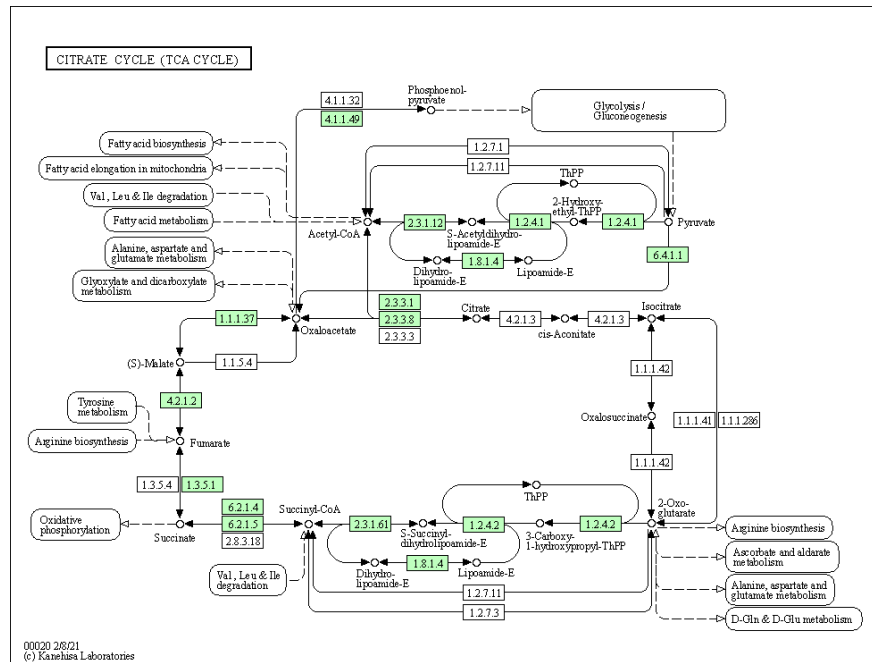


Figure 4.7 Annotated Citrate Acid Cycle pathway. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

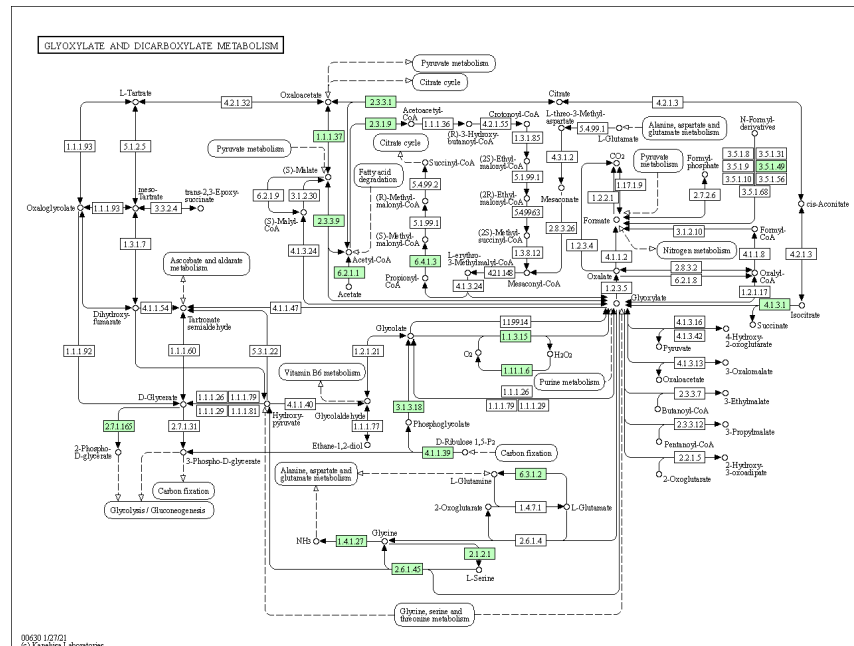


Figure 4.8 Annotated Glyoxylate and Dicarboxylate Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

RGd-1 was found to have nearly complete metabolic pathways for fatty acid elongation using acetyl-CoA in the mitochondrion or malonyl-CoA in the cytoplasm. The ability to switch between two different initial substrates may confer an advantage for fatty acid and neutral lipid biosynthesis. In September 2013, Witch Creek had a total nitrogen (TN) and dissolved inorganic carbon (DIC) concentrations of 125 ppb and 13.5 ppm, respectively. The ability to accumulate high lipid concentrations provides an ample storage of carbon and electrons that can be recycled and used to generate ATP during low nutrient conditions.

RGd-1 had a nearly complete fatty acid degradation pathway. Occurring in the mitochondria, fatty acid catabolism is important for recycling carbon to generate acetyl-CoA for the TCA cycle. Fatty acid oxidation involves multiple cycles of oxidation causing cleavage at the $C_{\alpha}-C_{\beta}$ at the β -carbon. Each catabolic cycle generates 2 carbon molecules in the form of acetyl-CoA and releases 4 electrons generated as NADH and $FADH_2$ driving the electron transport pathway and oxidative phosphorylation for the production of ATP.²¹²

Glycerolipid metabolism is important for the biosynthesis of neutral lipids such as TAGs. Glycerolipids are formed from a precursor, phosphatidic acid. Glycerolkinase catalyzes the phosphorylation of glycerol, forming glycerol-3-phosphate, the glycolysis intermediate and building block for glycerolipid synthesis. Acylated at the 1 and 2, positions, producing glycerol-3-phosphate acyltransferase. Acyltransferase adds the first acyl chain, forming monoacylglycerides (MAG), with the concomitant reduction of the backbone with NADPH and catalyzed by acyldihydroxyacetone phosphate reductase. The addition of acyl groups leads to the formation diacylglycerides (DAG) and triacylglycerides (TAG).²¹²

Figure 4.9 Annotated fatty acid metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

The RGd-1 genome had multiple copies of Ricin B lectin domain that is normally identified in castor beans and the ergot fungus, *Claviceps*,²¹³ with e-values ranging from 0.001-7.848e⁻⁷. Ricin has been found to have antifungal activity and was co-expressed with the Palmitic acid-specific elongase gene in the diatom, *Chaetoceros gracilis*, and may be transesterified to the glycerol backbone.²¹³

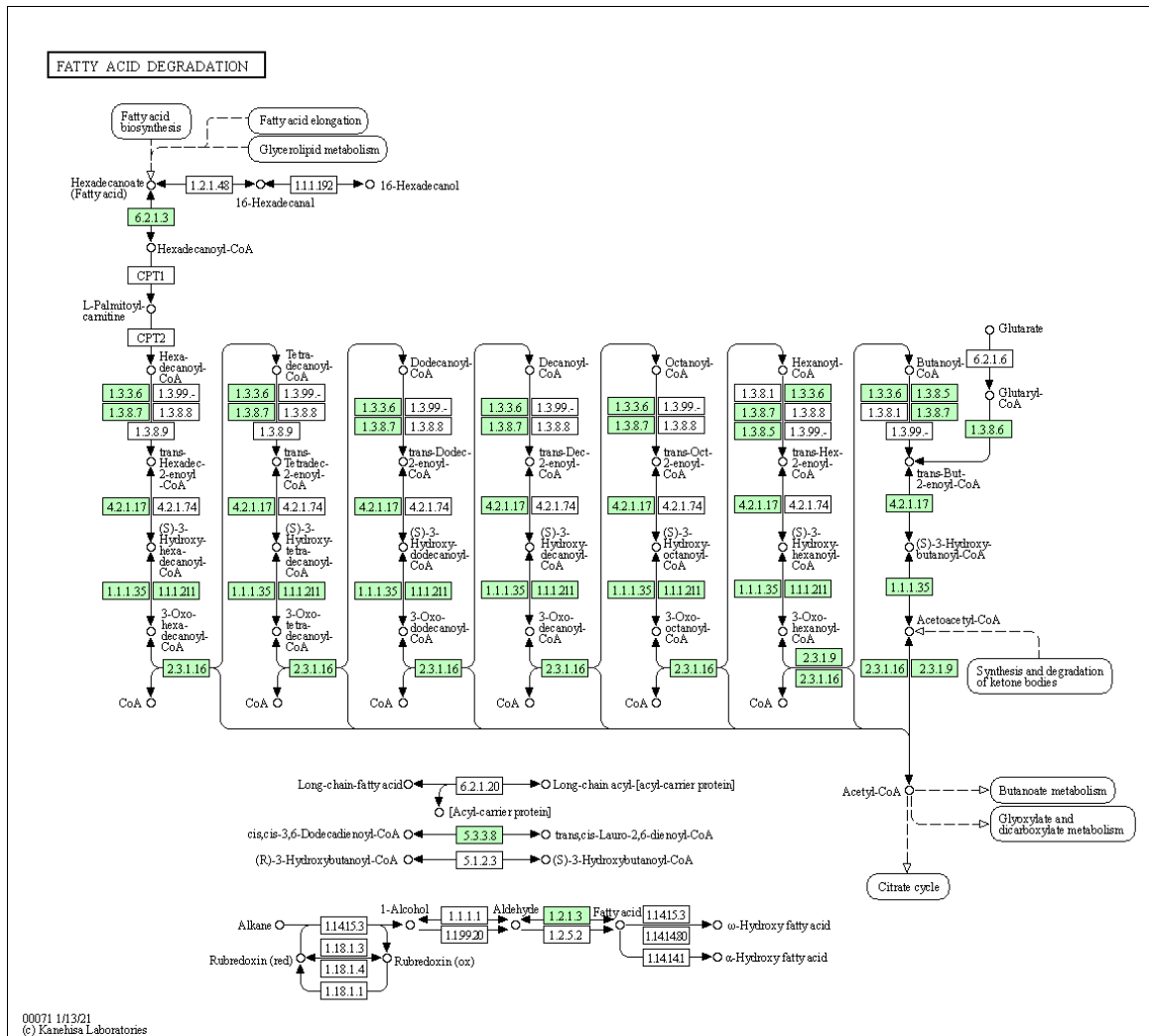


Figure 4.10 Annotated Fatty Acid Degradation Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

A simple experiment may provide the answer to this hypothesis: split the RGd-1 culture into 2 “lines”. One line will be transferred on strict timelines every two weeks. The other line will be transferred less often. It is hypothesized that the older culture will accumulate ricinoleic acid or a derivative if it is produced by RGd-1. Older cultures may have slower growth rates and distorted cell morphologies. The production of ricin, in the form of ricinoleic acid, may potentially contribute to the high concentrations of lipids produced by RGd-1. There is supporting evidence that RGd-1 may produce ricinoleic acid, an unsaturated C18 fatty acid in Moll et al., 2014, where C18:1-3 unsaturated fatty acids were the third most abundant in the RGd-1 FAME data. I hypothesize that the production of ricin may confer an anti-predatory or other fitness advantage in Witch Creek where there is a constant flow of creek water that can diffuse ricin away from the cells. Potentially, ricin may have at least two benefits, where it contributes to the pool of long-chain fatty acids and protects them from predation. However, in batch culture, the production of ricin may become toxic to the cells, especially for older cultures.

Figure 4.11 Annotated Glycerolipid Metabolism. The green boxes represent genes that are present in the RGd-1 genome. The annotated pathway was produced using KeggMapper.¹⁹⁰

Phycosphere Bacteria

Observations suggest RGd-1 grows in close association with bacteria, becoming unviable in antibiotic treated cultures. Nine different bacterial taxa were identified in the RGd-1 culture. Two possible functions, among many, for RGd-1 phycosphere bacteria are (1) aiding in arsenic reduction and uptake and (2) providing siderophores that may improve iron bioavailability in alkaline environments. The three most abundant classified taxa were *Agrobacterium* 40.9%, *Geobacter* 11.1%, and *Riemerella* 6.2%. Each of these genera is known to produce siderophores, including catechols (*Rhizobium* and *Afipia*)²¹⁴ and hydroxamates (*Brevundimonas*). Interestingly, both *Rhizobium* and *Afipia* contain many species of nitrogen-fixing bacteria often associated with plant roots. *Rhizobium*, *Brevundimonas*, and *Afipia* are known for siderophore production, which may be beneficial for diatom growth especially in alkaline environments such as the one from which RGd-1 was isolated.²¹⁵ In all, the seven populations observed here may be able to produce siderophores as other strains within those genera have consistently been found to produce siderophores. Specifically, members of the genera *Geobacter*, *Agrobacterium*, *Brevundimonas*, *Magnetococcus*, *Niastella*, and *Riemerella* have been observed to produce siderophores or have the genetic potential to produce siderophores.^{214, 216-221} Fifty-six percent of the bacteria observed here were unable to be classified using the RDB database for identification within CLARK, indicating taxonomic divergence from the reference organisms (Table 4.7).

As part of RGd-1 PacBio C4P6 sequencing, a bacterial genome was also assembled, and may represent a potential symbiont.²²² The bacterial genome had a 99% identity and 44% query coverage with *Brevundimonas* sp. within the Caulobacteraceae family. Due to the low query coverage, *Brevundimonas* sp., identified as KM-427, may represent a new species. KM-427 was found to have genes for siderophore biosynthesis (enterobactins and brucebactins), ferrichrome

iron receptors, Fe outer membrane receptor proteins, ferrochelatase, ferric uptake regulation protein (FUR), ferric iron ABC transporter and ferric enterobactin receptor. Further, this

Brevundimonas sp., KM-427 was also found to have genes for arsenic resistance and reduction:

ArsH (arsenic resistance protein), ACR3 (arsenic resistance protein), ArsR (transcription regulatory protein), arsenate reductase glutaredoxin and arsenical resistance operon repressor.

These genes may be important for survival in Witch Creek, a low-iron and high arsenic environment.²²³

Table 4.8 Identification for 16S amplified sequencing in the RGd-1 culture. Organisms were identified using the 16S RDB Database within CLARK.^{199, 200} Each organism was calculated for the percentage of all of the categories below (the 9 genera identified and the unknowns) and the percentage classified (the 9 genera excluding the unknowns).

Name	Tax ID	Count	Percentage of All (%)	Percentage of Classified (%)
<i>Agrobacterium</i>	373	28924	17.82	40.89
<i>Geobacter</i>	351604	7819	4.82	11.05
<i>Riemerella</i>	34085	4390	2.71	6.20
<i>Niastella</i>	354356	3899	2.40	5.51
<i>Magnetococcus</i>	1124597	3236	1.99	4.57
<i>Halothiobacillus</i>	927	3236	1.85	4.24
<i>Brevundimonas</i>	74313	2854	1.76	4.03
<i>Agrobacterium</i>	H13	2545	1.57	3.60
<i>Sulfuricurvum</i>	148813	1626	1.00	2.30
UNKNOWN	N/A	91472	56.38	N/A

Discussion

Genome Observations

Diatoms have immense, untapped biochemical potential due to their strong CO₂ utilization activity and ability to produce unique compounds, but with only five sequenced diatom genomes, our knowledge of these organisms is very limited. The five publicly available diatom genome assemblies are marine in origin, with no freshwater diatom genomes available.

The addition of the genome from the extremophilic, freshwater strain, RGd-1, will provide further insight about diatom taxonomy, physiology, metabolism, and evolution.

The RGd-1 draft genome has 520 contigs and 57.8% complete single-copy BUSCOs using the eukaryota lineage. When corrected for the total amount of complete orthologs that *P. tricornutum* has, RGd-1 only has 83.8% complete orthologs and 11.2% missing. This indicates that while the BUSCO scores can inform the gene capture rate based on the quality of the genome assembly, it is dependent on the gene models that were used to create each lineage and the diatom lineage is evolutionarily distant from other organisms included in presented comparisons. Since there is a 17% difference between the number of complete orthologs found in the RGd-1 and *P. tricornutum* assemblies, this indicates that while the RGd-1 genome assembly was able to capture a large proportion of orthologs, there is room for genome assembly improvement to capture more gene space.

RGd-1 was the most similar to *T. pseudonana* on a nucleotide level. This is surprising because RGd-1 and *T. pseudonana* belong to different lineages within class Bacillariophyceae, the pennates, and the centrics, which describes the diatom cell morphology. Considering that each of the diatoms with publicly available genome assemblies is marine in origin, it is reasonable that there would be a divergence between freshwater and marine diatoms. RGd-1 and *T. pseudonana* may have convergent shared functions that cause them to have a greater sequence identity. Despite RGd-1 being the most evolutionarily distant on a nucleotide level from *P. tricornutum*, when translated to proteins, there is more in common between these two diatoms indicating that there likely is a shared function for many proteins.

The *de novo* transcriptome assembly had better assembly statistics and BUSCO scores than the reference-guided transcriptome assembly which is consistent with the idea that any

errors in the genome assembly will be carried into the transcriptome assembly. In *de novo* mode, there is no reference bias and the reads can align optimally.

Metabolic Observations

RGd-1 was found to have nearly complete central carbon metabolism pathways. Specifically, RGd-1 showed a nearly complete glyoxylate pathway, where the TCA cycle is diverted from isocitrate to malate, conserving two carbons that would have been lost as decarboxylation steps in the TCA cycle (Figure 4.12). This is a potential contribution as to why RGd-1 can accumulate such high concentrations of palmitoleic acid (C16:1), palmitic acid (C16:0), C18:1-3 and eicosapentaenoic acid (C20:5).⁴⁹ While *P. tricornutum* has been shown to use the glyoxylate shunt,²²⁴ to the best of our knowledge, no one has contextualized the glyoxylate shunt in terms of its potential role in improving lipid accumulation and possibly for biofuel production. Further analysis using differential gene expression may shed important light on the level of expression of genes in the glyoxylate shunt pathway under different lipid accumulating conditions such as carbon deplete and replete growth conditions.

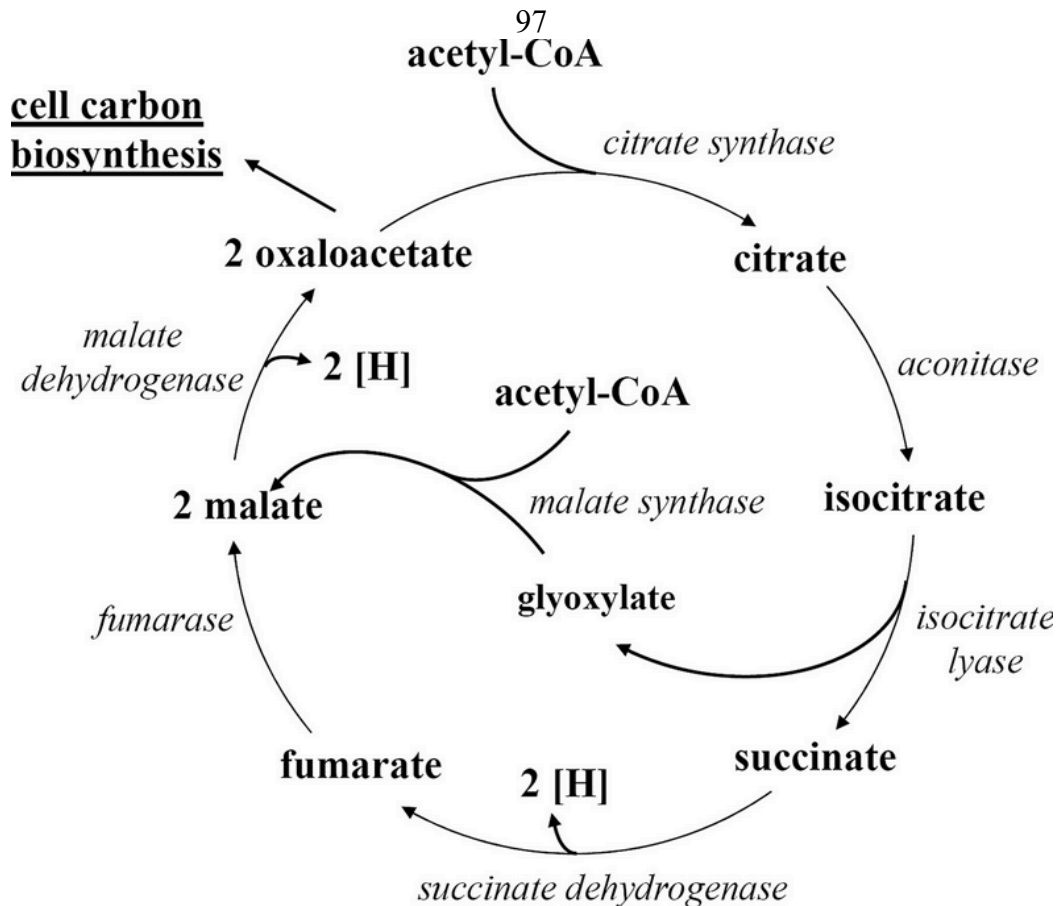


Figure 4.12 The citric acid cycle with the glyoxylate and dicarboxylate pathway which diverts isocitrate to malate.²²⁵ The two enzymes, isocitrate lyase and malate synthase modify the citric acid cycle avoiding two decarboxylation steps resulting in the formation of malate from 2 molecules of acetyl-CoA.²²⁵

The organisms inhabiting Witch Creek experience a wide range of conditions throughout the year. The shallow stream experiences temperature shifts from the hot effluent channels that feed into the stream, while being impacted by high levels of snowfall during the winter months. In addition to the temperature shifts, the creek receives high concentrations of silica and arsenic from the effluent channels and is generally limited in nitrogen and inorganic carbon. The ability to accumulate high concentrations of lipids via different metabolic pathways offers a greater potential for success during the bleaker months of Yellowstone National Park. Through β -

oxidation of long-chain fatty acids, RGd-1 can recycle carbon and electrons towards maintenance energy during the harsher months when there are fewer nutrients in an already nutrient-limited environment and the temperature is sub-optimal for diatom growth.

Bacterial Cohabitants

RGd-1 grows well as a unialgal culture. The effects of the bacterial co-habitants are currently unknown. The relationships between bacteria and algae can be complex. *Phaeobacter galliensis* has been found to secrete different compounds to either stimulate growth (auxins) of *Emiliania huxleyi* or induce cell lysis with algaecides for older cells, indicating that *P. galliensis* can switch from symbiotic to a parasitic relationship with *E. huxleyi* depending on the stage of growth of the cells.²²⁶ While most algal cultures, including diatoms, have been grown as axenic, or unialgal cultures, there has been a recent paradigm-shift towards growing algae in mixed cultures to provide more stable growth and productivity.^{148, 227} Studying the mechanisms of cross-feeding between phototrophs and heterotrophs, a subject about which relatively little is known, may provide fundamental insight into diatom growth and critical microbial community interactions. Further, it is important to devise diatom growth strategies that provide the best opportunities for improving the rate and extent of biofuel precursors and high-value products. Understanding phototroph-heterotroph interactions would improve the ability to implement diatom growth strategies that exploit these interactions.

The 16S amplified community analysis was performed post-antibiotic treatments after it was discovered that *Brevundimonas* sp. was present and the majority of the original PacBio reads were prokaryotic. Given the distribution of the bacteria in the RGd-1 culture it is surprising that the dominant bacterium, *Agrobacterium*, was not sequenced and assembled as part of the PacBio sequencing project. Following this discovery, aggressive antibiotic treatments were employed

and MiSeq 16S sequencing was used to determine what bacteria were potentially still present in the RGd-1 culture. Therefore, these treatments likely decreased *Brevundimonas* sp. concentrations and the overall bacterial community would have shifted as a result.

The association of bacteria with diatoms within a phycosphere can be thought of as a diatom microbiome. This is an important concept in the study of algae and is often overlooked. While some bacteria associated with diatoms have been identified, their specific interactions are much less well defined,⁴⁴ with the literature being particularly scarce for non-marine diatoms. Bacteria attached to diatom frustules have been found to exchange substances such as vitamins, indole, and organic carbon compounds with their associated diatom and to receive protection from predators (Figure 4.13).¹⁶¹⁻¹⁶⁴ Attached bacteria have been hypothesized to promote organic matter degradation, which may or may not result in CO₂. I hypothesize that this may be one explanation for the overall lower TAG accumulation that is observed in mixed communities compared to unialgal cultures.¹⁰²

When carbon is passed between bacteria to algae increasing community biomass, sedimentation occurs providing organic carbon to benthic communities.^{44, 159} However, it is currently not known whether freshwater diatoms and bacteria form symbiotic relationships to overcome the diatom's inability to produce iron-scavenging siderophores. These inter-domain interactions are anticipated to be especially important in high-pH environments, characterized by very low iron concentrations, such as Witch Creek, the environment from which RGd-1 was isolated.

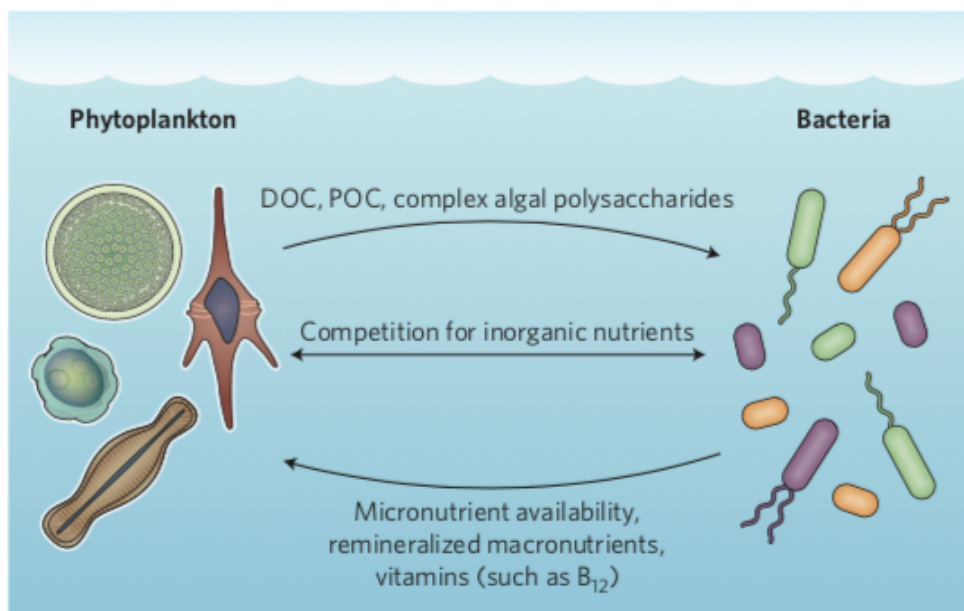


Figure 4.13 Potential mechanisms of symbiosis between marine diatoms and bacteria. Phytoplankton, such as diatoms may provide dissolved organic carbon (DOC), particulate organic carbon (POC), and other complex algal polysaccharides. The bacteria may supply micronutrients, macronutrients, and vitamins such as B₁₂.^{159,161-164}

Multiple bacteria identified in the RGd-1 culture (Table 4.8) are closely related to organisms that have been shown to produce siderophores. Various *Agrobacterium* strains have been found to produce hydroxamates or agrobactin.^{214, 217} *Brevundimonas diminuta* has been found to produce siderophores within the rhizosphere of *Oryza sativa* (rice) as determined by the CAS assay.^{228, 229} *Niastella* is a member of the phylum Bacteroidetes and multiple species within the *Niastella* genus and related strain, *Arachidicoccus rhizosphaerae*, have been found to produce plant promoting properties such as indole-3-acetic acid, siderophores (as detected by the CAS assay), and NH₃.^{220, 229} *Sulfuricurvum* also has a complete siderophore biosynthesis pathway. Furthermore, *G. uraniireducens* is capable of dissimilatory iron reduction (DIRB),^{230, 231} which, in addition to siderophore production, may provide a unique source of iron availability by reducing Fe(III) to Fe(II), a more biologically available form. It is surprising that *Geobacter*

constituted 4.82% of the total bacteria identified in the RGd-1 culture given its strict anaerobic growth requirements in a highly oxic environment.²²⁰ One explanation may be that there are microenvironments on or within the RGd-1 frustules that are shielded from oxygen. Witch Creek has a pH of 9.3 and is very low in iron. It is possible that in this environment, symbioses may form between diatoms and phycosphere bacteria so that the bacteria chelate any iron that may be available in Witch Creek and bring the iron in close proximity to the diatom for cell surface reduction and/or uptake.

Conclusions

Diatom strain RGd-1 is a novel diatom that is genetically divergent from the marine diatoms with publicly available sequenced genomes. A comparison of two genome assemblies showed that although one contained longer read lengths, it had worse genome statistics and BUSCO scores. Therefore, RGd-1 v. 1.0 was used for all downstream analyses. The *de novo* transcriptome assembly had better assembly statistics and BUSCO scores compared to the reference-guided transcriptome assembly and was used for the RGd-1 v. 1.0 genome annotation. The RGd-1 genome annotation revealed evidence that RGd-1 has the glyoxylate shunt pathway that could be used as a carbon conservation strategy; a mechanism that may contribute to observations that RGd-1 can accumulate very high concentrations of neutral lipids. The glyoxylate shunt has been identified in *T. pseudonana*, *P. tricornutum*²³², and *C. fusiformis*.²³³ Next-generation sequencing was used to identify bacteria found in the RGd-1 cultures. Important dynamics may be occurring between RGd-1 and its phycosphere bacterial community. While cross-feeding of vitamins and nutrients between diatoms and bacteria have been documented, there may be other important mechanisms of symbiosis that, to date, are largely unexplored.

Nine bacterial community members were observed that may play key functions (e.g. iron transfer), and contribute to diatom health in low iron and high arsenic environments.

Algae are often claimed to be axenic in the literature. The inclusion of next-generation sequencing has increased the sensitivity to determine bacterial contaminants, cohabitants, and potential phycosphere organisms. Using GC content and codon bias of observed sequences, we were able to assemble the genome of a new species of *Brevundimonas*, which was shown to have the genetic potential to reduce and assimilate heavy metals such as arsenic. Future work is needed to elucidate the stability and various functions of the RGd-1 phycosphere community.

103
CHAPTER FIVE

GENOME SEQUENCE FOR AN NOVEL BREVUNDIMONAS STRAIN

Contribution of Authors and Co-Authors

Manuscript in Chapter 5

Karen M. Moll, Nico Devitt, Thiruvarangan Ramaraj, Joann Mudge, Brent M. Peyton

Author: Karen M. Moll

Contributions: Performed the analysis, wrote the paper

Co-Author: Thiruvarangan Ramaraj

Contributions: Provided mentorship and guidance for the analyses

Co-Author: Nicholas P. Devitt

Contributions: Provided mentorship and guidance for the analyses

Co-Author: Joann Mudge

Contributions: Provided mentorship and guidance for the analyses and paper writing

Co-Author: Brent M. Peyton

Contributions: Provided mentorship and guidance for the analyses and paper writing

Manuscript Information

Karen M. Moll, Nico Devitt, Thiruvarangan Ramaraj, Joann Mudge, Brent M. Peyton
Microbiology Resource Announcements

Status of Manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☐ Published in a peer-reviewed journal

105
Abstract

Brevundimonas sp. strain KM-427 was found to potentially be a phycosphere bacterium associated with the extremophilic diatom, RGd-1. Here, we present the complete genome assembly and annotation for *Brevundimonas* sp. strain KM-427 that was sequenced as part of a PacBio sequencing project for the diatom, RGd-1. This genome provides insight into the *Brevundimonas* genus.

Announcement

Diatom strain, RGd-1, was isolated from an alkaline stream in Yellowstone National Park, WY, USA.⁴⁹ Genomic DNA from the unialgal culture was extracted using a phenol-chloroform extraction to isolate high-molecular-weight DNA. To sequence this DNA, 2 µg of purified high-molecular-weight DNA was prepared. The BluePippen size selection at 10 Kb lengths was used for sample preparation prior to sequencing.²³⁴ Three SMRTcells were run with C4P6 chemistry (6-hour movies) to improve read length and coverage at The National Center for Genome Resources (Santa Fe, NM). The resulting sequence was approximately 5% diatom and 95% bacterial, eventually determined to belong to genus, *Brevundimonas*.

Brevundimonas sp. was sequenced and assembled completely into one contig with a 3.1 Mb genome length and 68.85% GC content (Table 5.1). The reads were assembled using Canu and two genomes were assembled, diatom strain RGd-1 and *Brevundimonas* strain, KM-427. The *Brevundimonas* sp. genome was annotated using Prokka (version 1.12) and separately with RAST.²³⁵⁻²³⁷ KM-427 had 99% identity and 44% query coverage with *Brevundimonas* sp. within the Caulobacteraceae family. Due to the low query coverage, *Brevundimonas* sp., KM-427, may represent a new species.

Table 5.1 Genome assembly statistics for *Brevundimonas* sp., strain KM-427. The genome was assembled with Canu as part of an RGd-1 PacBio sequencing project.¹⁷⁵

Assembly Statistic	KM-427
Number of contigs	1
Genome length	3,090,090
Contig N50	3,090,090
Maximum contig size	3,090,090
Minimum contig size	3,090,090
GC content (%)	68.85

A total of 70.9% benchmarking universal single-copy orthologs (BUSCOs) were identified, 0.7% fragmented, and 28.4% missing BUSCOs within the bacteria lineage (Table 5.2). Since KM-427 was assembled entirely into one contig, the missing BUSCOs may represent unique orthologs that were not present in the bacteria lineage.³¹

Table 5.2 Gene capture measured by BUSCO. A total of 148 BUSCOs were searched within the bacteria odb9 lineage.

BUSCO type	Number of BUSCOs found	% BUSCOs found
Complete-single copy	175	57.8%
Complete-duplicated	3	1.0%
Fragmented	20	6.6%
Missing	105	34%

The *Brevundimonas* genome annotation revealed genes involved with copper homeostasis, arsenic resistance, and cobalt, zinc, and cadmium resistance (Figure 5.1).^{236, 237} KM-427 was found to contain genes for arsenic resistance, reduction, and uptake. The arsenic resistance operon with the following genes identified; *arsH*, *acr3*, *arsR*, and *arsM* (Figure 5.1). Arsenic metabolism may be important for the survival and adaptation of *Brevundimonas* sp. strain KM-427 since its natural environment containing 300 ppb of arsenic.¹⁰¹

The KM-427 genome annotation revealed genes for producing siderophore-producing genes that are used for chelating iron in iron-limiting environments. TonB and Tol transport systems and hemin transport systems were identified that are known to be involved in ferric siderophore transport and ABC transporters, ferrous iron transport protein A, ferrichrome iron receptor, ferric uptake regulation protein, FUR, ferric iron ABC transporter ATP-binding protein, iron-uptake factor PiuB, ferrous iron transport protein B, ferric iron ABC transporter iron-binding protein, and ferric iron ABC transporter permease protein. In addition, KM-427 was found to contain enterobactin transferase and brucebactin synthetase genes. Further, 43% of the annotated genes are considered “hypothetical proteins”, and it is possible that more siderophore biosynthesis genes are within this fraction.

Siderophore biosynthesis may also be beneficial to the diatom, RGd-1 from which KM-427 was identified and isolated. Previous studies have shown the ability of diatoms to utilize siderophores produced by bacteria.^{238, 239} In this case, siderophores may provide RGd-1 with bioavailable iron in an otherwise iron-limited environment.

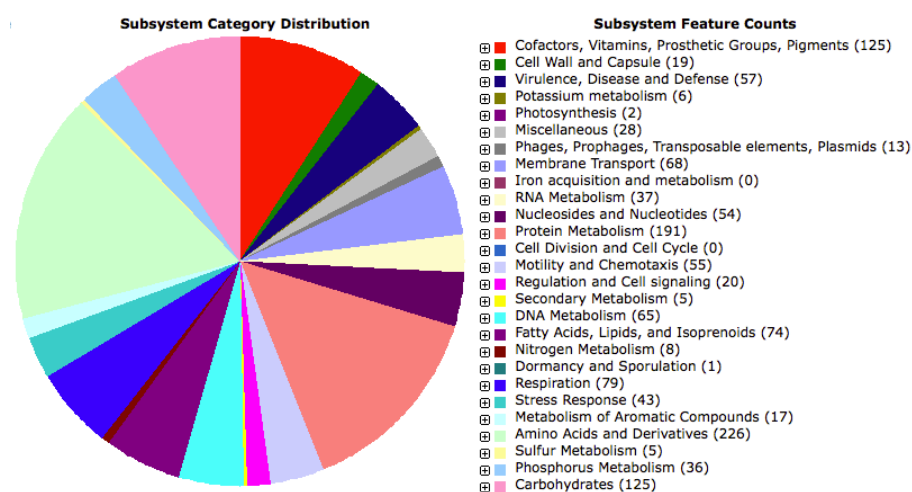


Figure 5.1 The *Brevundimonas* sp. strain, KM-427, genome annotation major features.

Data availability. The complete genome assembly is available under the following Genbank

BioSample accession number, SAMN12024285, strain 2588940, and BioProject PRJNA548375.

109
CHAPTER SIX

SUMMARY AND FUTURE DIRECTIONS

Synopsis

This dissertation focused on alkaliphic algae for biofuel applications. Algae have a strong potential in the renewable energy sector, however more research needs to be performed to bring the price at the pump equivalent to petroleum-based fuels. The combustion of petroleum-based fuels adds CO₂ to the atmosphere and is ultimately environmentally polluting and the combustion of petroleum-based fuels adds CO₂ into the atmosphere.^{52, 240} As a near carbon neutral biotechnology, algae provide a more environmentally sustainable source of fuels.

A type of algae that forms distinctive silica shells, diatoms are critical to ecosystem health at a global scale and have the capacity to buffer climate change. Producing at least 25% of Earth's atmospheric oxygen, diatoms also fix ~25-45% of global CO₂ directly mitigating the major driver behind climate change.¹² However, diatoms are not immune to the effects of climate change, manifested in aquatic systems as rapid temperature fluctuations and acidification. As a result, a drastic global decline in diatoms has recently been observed that will directly reduce oxygen contributions and carbon fixation, further compounding climate change.²⁴¹ Such compounding factors serve to accelerate climate change, and with very limited time to correct, now is the critical moment to truly understand diatom ecology, including the relationships with bacterial symbionts starting at the genomic-level. These relationships are vital to diatom productivity and survival, and are thus critical to oxygen production and carbon fixation. Without a fundamental understanding of these symbiotic relationships, diatoms could succumb to climate change faster than they can buffer its effects. While some work has studied symbiotic

relationships of marine diatoms, little is known about freshwater diatom symbiotic relationships, despite their often dominant presence in these systems.

Strain Selection

Algal strain selection has been an important focus for algal biodiesel research. We can select strains from environments with qualities most similar to those used in outdoor raceway ponds. Halophiles have been selected due to their salt tolerance as salinity increases with evaporative loss. In addition, conditions like high salinity or alkalinity can limit colonization by competing microorganisms. Here, we focused on isolating alkaliphilic strains found to be most productive for biofuels based on growth rate, dry cell weight, and lipid accumulating abilities.

Algal lipid concentration is influenced by several factors, including carbon availability. Some algae strains have been shown to increase their lipid concentrations when sodium bicarbonate is added just prior to nitrate depletion. However, not all strains respond to the sodium bicarbonate addition, and further work is needed to elucidate why some strains respond and some do not. Diatom strain, RGd-1, accumulated approximately two times the lipid concentrations compared when given sodium bicarbonate. Because RGd-1 was able to accumulate 70-80% FAMES w/w for ash free dry weight, interest was generated as to why it was able to accumulate such high concentrations of lipids.⁴⁹ The first step toward understanding its lipid metabolism was to sequence, assemble and annotate the RGd-1 genome. An annotated genome facilitates the identification of potentially novel genes or potential pathways that enable RGd-1 to accumulate very high lipid concentrations.

Using the RGd-1 genome assembly and annotation, it was possible to determine metabolic pathways used by RGd-1. RGd-1 had complete lipid synthesis and degradation

pathways and near-complete carbon metabolism pathways such as the TCA cycle, carbon fixation and uses the glyoxylate shunt within the TCA cycle. The glyoxylate shunt is of particular interest because it is well known to be a carbon conserving pathway that avoids two decarboxylations that occur in the full TCA cycle between isocitrate to α -ketoglutarate and α -ketoglutarate to succinyl-CoA with the enzymes, isocitrate dehydrogenase and oxalosuccinate decarboxylase, respectively. Use of the glyoxylate shunt may be one mechanism that allows RGd-1 to accumulate high concentrations of lipids.

In carbon-limited environments, like alkaline systems, it may be advantageous for a microorganism such as RGd-1 to develop lipid accumulating and carbon-conservation strategies for survival and maintenance. At alkaline pH, there is a greater flux of atmospheric CO₂ into the stream that will enrich the DIC concentration, allowing for greater access to carbon. However, any competing microorganisms that grow at a faster rate, such as cyanobacteria, will be able to utilize these resources leaving Witch Creek in a regular state of nutrient limitation. The ability for RGd-1 to utilize the glyoxylate cycle, use both acetyl-CoA or malonyl-CoA as starting substrates for fatty acid biosynthesis, and potentially use ricin as a fatty acid source may all contribute to high fatty lipid concentrations. Having a collection of lipid accumulating and carbon conserving strategies increase the potential for RGd-1's success.

Future Directions

One phycosphere inhabitant, *Brevundimonas* sp., has the genetic capacity to produce siderophores, making iron bioavailable to RGd-1. Two additional taxa, previously shown to produce siderophores,²²⁹ were identified in association with RGd-1. Sequencing has confirmed an additional six taxa living in association with RGd-1.⁴⁹ The relationship between RGd-1 and

these bacteria, particularly *Brevundimonas* sp., is an ideal focus due to potentially intertwined metabolite exchange. This relationship likely facilitates the survival of RGd-1 in a high pH (9.3), arsenic-rich (300 ppb), and iron poor environment. Exploring this relationship through CRISPR/CAS modifications will provide critical insights into the importance of symbiosis to freshwater diatom resilience. Using CRISPR/CAS as an antibiotic to systematically target the phycosphere bacteria to determine their effects on RGd-1 growth and lipid accumulation. By focusing on the removal of each phycosphere bacteria, it will be possible to gain a better understanding of inter-domain iron transfer.

Despite being vital to diatom health, the role of bacterial symbionts is often overlooked. Residing in the phycosphere, the outer sheath encapsulating a diatom, bacteria consume a nearly limitless carbon supply while occupying a niche virtually free of predators.⁴² In exchange, bacteria provide diatoms with vitamins like B₁₂, growth factors, siderophores and antimicrobials.^{159,161-164} Together, each organism is more efficient at its ecosystem function; bacteria increase organic matter degradation producing CO₂ and freeing carbon to lower trophic levels while diatom photosynthesis sequesters CO₂. However, CO₂ sequestration can dramatically elevate pH, limiting nutrient bioavailability. It is here that bacterial may symbiosis play one of its most critical roles, providing nutrients that are otherwise inaccessible. When pH ≥ 8, iron is precipitated in biologically unavailable forms, such as Fe(OH)₃. A strategy implemented by many bacteria in iron-limited environments is to produce siderophores, low molecular weight (0.5-1.5 kDa) Fe(III) chelating agents that establish multiple (normally hexavalent) bonds with iron solubilizing it, making it biologically available. While diatoms do not produce siderophores, they efficiently scavenge iron from siderophores produced by bacteria.

Although diatoms are unable to produce siderophores themselves, it has been shown that they can utilize iron complexed with bacterial siderophores^{57, 242} and that diatoms interact with siderophore-producing bacteria.⁴⁴ For instance, previous studies have found that diatoms can utilize exogenously produced siderophores such as desferrioxamine B and E, and ferrioxamines D and E when grown in iron deficient media.^{238, 239} At the diatom cell surface, iron is thought to be reduced by membrane-bound ferrireductase and taken up by iron permease.²⁴³ Weak siderophore affinity for Fe(II) allows dissociation of iron from the siderophore complex, enabling iron entry into the cell. Further, there is genetic evidence for the direct uptake of siderophores by diatoms. The ferrichrome binding protein (FBP) gene in *P. tricornutum* is associated with siderophore uptake and works in concert with ferric reductase (FRE) to sequester and convert Fe to biologically available forms. Kazamia et al. 2018 found that *P. tricornutum* endocytized the siderophore-Fe complex, especially under iron-limiting conditions when iron starvation-induced protein 1 (ISIP1) was highly expressed.²⁴⁴

Lipids are also important for plant-microbe interactions. Phospholipids, sphingolipids, glycolipids, and sterol lipids are important to establish plant microbe interactions. Some lipids such as diacylglycerol, phosphaditic acid, and free fatty acids are signaling molecules for plants and microbes.²⁴⁵ These types of molecules may be important in low-iron environments. Here the diatom may detect low iron, followed by an increased expression of ISIP as seen in *P. tricornutum*. Potentially, the diatom may produce a signaling molecule in the form of a lipid that may tell an associated bacteria to produce siderophores. Some bacteria produce TAGs, wax esters or polyhydroxyalkanoates (PHAs) that can be use ultimately used for biodiesel.²⁴⁶ It is possible that the bacteria associated with RGd-1 produce biodiesel-producing TAGs or free fatty acids that become part of the RGd-1 FAME pool when transesterified. More work is needed to

determine whether the bacteria associated with RGd-1 have the genetic and physiological ability to produce signaling and/or biodiesel containing lipids.

Future Work

1. The glyoxylate shunt will be examined more closely for its potential role in the very high lipid accumulating abilities of RGd-1.
2. Determine the metabolic potential for RGd-1 to make Ricin.

Closing

The work presented here demonstrates the novelty of the extremophilic, freshwater, diatom strain RGd-1 from a genomics perspective. RGd-1 is divergent from the five publicly available marine diatom genome assemblies, which will provide insight into diatom ecology and evolution. Further, it is possible to gain insight into key genes that have allowed RGd-1 to become adapted to its alkaliphilic environment containing high concentrations of heavy metals such as arsenic. With additional omics approaches, combined with physiological data, it will be possible to obtain a deep understanding of what makes RGd-1 unique. Lastly, it is important to consider not only one organism but of how multiple organisms function together in their environment. RGd-1 and *Brevundimonas* sp. may have a dependent relationship and it is important to determine what factors drive these types of relationship. The complex dynamics between the co-habiting bacteria and diatoms may operate so efficiently as to function as a single organism.

References

1. *Global Transport Scenarios 2050*. 2011; Available from: https://www.worldenergy.org/wp-content/uploads/2012/09/wec_transport_scenarios_2050.pdf.
2. *Oil Market Report*. 11 April 2019; Available from: <https://www.iea.org/oilmarketreport/omrpublic/>.
3. Sheehan, J., Dunahay, T. Benemann, J. & Roessler, P., *A look back at the U.S. Department of Energy's Aquatic Species Program- biodiesel from algae*. National Renewable Energy Laboratory. 1998.
4. Fields, M.W., et al., *Sources and resources: importance of nutrients, resource allocation, and ecology in microalgal cultivation for lipid accumulation*. Appl Microbiol Biotechnol, 2014. **98**(11): p. 4805-16.
5. Chisti, Y., *Biodiesel from microalgae*. Biotechnol Adv, 2007. **25**(3): p. 294-306.
6. Georgianna, D.R. and S.P. Mayfield, *Exploiting diversity and synthetic biology for the production of algal biofuels*. Nature, 2012. **488**(7411): p. 329-335.
7. Courchesne, N.M., et al., *Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches*. J Biotechnol, 2009. **141**(1-2): p. 31-41.
8. Greenwell, H.C., et al., *Placing microalgae on the biofuels priority list: a review of the technological challenges*. J R Soc Interface, 2010. **7**(46): p. 703-26.
9. Razeghifard, R., *Algal biofuels*. Photosynth Res, 2013. **117**(1-3): p. 207-19.
10. Trentacoste, E.M., et al., *Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth*. Proc Natl Acad Sci U S A, 2013. **110**(49): p. 19748-53.
11. d'Ippolito, G., et al., *Potential of lipid metabolism in marine diatoms for biofuel production*. Biotechnol Biofuels, 2015. **8**: p. 28.
12. Cavicchioli, R., et al., *Scientists' warning to humanity: microorganisms and climate change*. Nat Rev Microbiol, 2019.
13. Mock, T., et al., *Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses*. Proceedings of the National Academy of Sciences, 2008. **105**(5): p. 1579-1584.
14. Fleischer, K., et al., *Amazon forest response to CO₂ fertilization dependent on plant phosphorus acquisition*. Nature Geoscience, 2019. **12**(9): p. 736-741.
15. Markou, G. and E. Nerantzis, *Microalgae for high-value compounds and biofuels production: a review with focus on cultivation under stress conditions*. Biotechnol Adv, 2013. **31**(8): p. 1532-42.
16. Yandell, M. and D. Ence, *A beginner's guide to eukaryotic genome annotation*. Nat Rev Genet, 2012. **13**(5): p. 329-42.
17. Metzker, M.L., *Sequencing technologies - the next generation*. Nat Rev Genet, 2010. **11**(1): p. 31-46.
18. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies*. Nat Rev Genet, 2016. **17**(6): p. 333-51.
19. Ambardar, S., et al., *High Throughput Sequencing: An Overview of Sequencing Chemistry*. Indian J Microbiol, 2016. **56**(4): p. 394-404.

20. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 2005. **437**.
21. Lang, D., et al., *Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford Nanopore*. bioRxiv, 2020.
22. Hon, T., et al., *Highly accurate long-read HiFi sequencing data for five complex genomes*. bioRxiv, 2020.
23. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications*. Nat Rev Genet, 2020. **21**(10): p. 597-614.
24. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome*. Nat Biotechnol, 2019. **37**(10): p. 1155-1162.
25. Lam, E.T., et al., *Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly*. Nat Biotechnol, 2012. **30**(8): p. 771-6.
26. Pendleton, M., et al., *Assembly and diploid architecture of an individual human genome via single-molecule technologies*. Nat Methods, 2015. **12**(8): p. 780-6.
27. Bickhart, D.M., et al., *Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome*. Nat Genet, 2017. **49**(4): p. 643-650.
28. Moll, K.M., et al., *Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula*. BMC Genomics, 2017. **18**(1): p. 578.
29. Putnam, N.H., et al., *Chromosome-scale shotgun assembly using an in vitro method for long-range linkage*. Genome Res, 2016.
30. Simpson, J.T., et al., *ABYSS: a parallel assembler for short read sequence data*. Genome Research, 2009. **19**(6): p. 1117-1123.
31. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs*. Bioinformatics, 2015. **31**(19): p. 3210-2.
32. Vurture, G.W., et al., *GenomeScope: fast reference-free genome profiling from short reads*. Bioinformatics, 2017. **33**(14): p. 2202-2204.
33. Marcais, G. and C. Kingsford, *A fast, lock-free approach for efficient parallel counting of occurrences of k-mers*. Bioinformatics, 2011. **27**(6): p. 764-70.
34. Mapleson, D., et al., *KAT A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies*. Bioinformatics, 2017. **33**(4): p. 574-576.
35. Bowler, C., et al., *The Phaeodactylum genome reveals the evolutionary history of diatom genomes*. Nature, 2008. **456**(7219): p. 239-44.
36. Armbrust, E.V., et al., *The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism*. Science, 2004. **306**(5693): p. 79-86.
37. Traller, J.C., et al., *Genome and methylome of the oleaginous diatom Cyclotella cryptica reveal genetic flexibility toward a high lipid phenotype*. Biotechnol Biofuels, 2016. **9**: p. 258.
38. Helliwell, K.E., et al., *Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes*. Mol Biol Evol, 2011. **28**(10): p. 2921-33.
39. Bowler, C., A. Vardi, and A.E. Allen, *Oceanographic and biogeochemical insights from diatom genomes*. Ann Rev Mar Sci, 2010. **2**: p. 333-65.

40. Mock, T., et al., *Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus*. Nature, 2017. **541**(7638): p. 536-540.
41. Bell, T.A.S., et al., *Microbial community changes during a toxic cyanobacterial bloom in an alkaline Hungarian lake*. Antonie Van Leeuwenhoek, 2018.
42. Mitchell, W.B.a.R., *CHEMOTACTIC AND GROWTH RESPONSES OF MARINE BACTERIA TO ALGAL EXTRACELLULAR PRODUCTS*. Biological Bulletin, 1972. **143**(2): p. 265-277.
43. Lang, W.H.B.a.J.M., *Selective stimulation of marine bacteria by algal extracellular products*. Limnology and Oceanography, 1974. **19**(5): p. 833-839.
44. Amin, S.A., M.S. Parker, and E.V. Armbrust, *Interactions between diatoms and bacteria*. Microbiol Mol Biol Rev, 2012. **76**(3): p. 667-84.
45. Krohn-Molt, I., et al., *Insights into Microalga and Bacteria Interactions of Selected Phycosphere Biofilms Using Metagenomic, Transcriptomic, and Proteomic Approaches*. Front Microbiol, 2017. **8**: p. 1941.
46. Rolland, J.L., et al., *Quorum Sensing and Quorum Quenching in the Phycosphere of Phytoplankton: a Case of Chemical Interactions in Ecology*. J Chem Ecol, 2016. **42**(12): p. 1201-1211.
47. Sapp, M., et al., *Species-specific bacterial communities in the phycosphere of microalgae?* Microb Ecol, 2007. **53**(4): p. 683-99.
48. Wang, H., et al., *Effects of bacterial communities on biofuel-producing microalgae: stimulation, inhibition and harvesting*. Crit Rev Biotechnol, 2016. **36**(2): p. 341-52.
49. Moll, K.M., et al., *Combining multiple nutrient stresses and bicarbonate addition to promote lipid accumulation in the diatom RGd-1*. Algal Research, 2014. **5**: p. 7-15.
50. Zilber-Rosenberg, I. and E. Rosenberg, *Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution*. FEMS Microbiol Rev, 2008. **32**(5): p. 723-35.
51. Bhateria, R. and R. Dhaka, *Algae as biofuel*. Biofuels, 2015: p. 1-25.
52. Hill, J., et al., *Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels*. Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11206-10.
53. Markou, G., D. Vandamme, and K. Muylaert, *Microalgal and cyanobacterial cultivation: the supply of nutrients*. Water Res, 2014. **65**: p. 186-202.
54. Brennan, L. and P. Owende, *Biofuels from microalgae—a review of technologies for production, processing, and extractions of biofuels and co-products*. Renewable and sustainable energy reviews, 2010. **14**(2): p. 557-577.
55. Hu, Q., et al., *Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances*. Plant J, 2008. **54**(4): p. 621-39.
56. Schenk, P., et al., *Second Generation Biofuels: High-Efficiency Microalgae for Biodiesel Production*. BioEnergy Research, 2008. **1**(1): p. 20-43.
57. Amin, S.A., et al., *Photolysis of iron, siderophore chelates promotes bacterial, algal mutualism*. Proceedings of the National Academy of Sciences, 2009. **106**(40): p. 17071-17076.
58. Seckbach, J., *Algae and cyanobacteria in extreme environments*. 2007, The Netherlands: Springer.
59. Hildebrand, M., et al., *The place of diatoms in the biofuels industry*. Biofuels, 2012. **3**(2): p. 221-240.

60. Reinfelder, J.R., A.J. Milligan, and F.M. Morel, *The role of the C₄ pathway in carbon accumulation and fixation in a marine diatom*. Plant Physiol, 2004. **135**(4): p. 2106-11.
61. Roberts, K., et al., *Carbon acquisition by diatoms*. Photosynthesis Research, 2007. **93**(1): p. 79-88.
62. Giordano, M., J. Beardall, and J.A. Raven, *CO₂ CONCENTRATING MECHANISMS IN ALGAE: Mechanisms, Environmental Modulation, and Evolution*. Annual Review of Plant Biology, 2005. **56**(1): p. 99-131.
63. Kaplan, A. and L. Reinhold, *CO₂ CONCENTRATING MECHANISMS IN PHOTOSYNTHETIC MICROORGANISMS*. Annual Review of Plant Physiology & Plant Molecular Biology, 1999. **50**(1): p. 539.
64. Somanchi, J.V.M.a.A., *How Do Algae Concentrate CO₂ to Increase the Efficiency of Photosynthetic Carbon Fixation?* Plant Physiology, 1999. **119**: p. 9-16.
65. Moroney, J.V. and R.A. Ynalvez, *Proposed carbon dioxide concentrating mechanism in Chlamydomonas reinhardtii*. Eukaryot Cell, 2007. **6**(8): p. 1251-9.
66. Radakovits, R., et al., *Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana*. Nat Commun, 2012. **3**: p. 686.
67. Jacob Valenzuela, A.M., Ross P Carlson, Robin Gerlach, Keith E Cooksey, Brent M Peyton and Matthew W Fields, *Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum*. Biotechnology for Biofuels, 2012. **5**(40): p. 1-17.
68. Bigelow, N., et al., *A Comprehensive GC-MS Sub-Microscale Assay for Fatty Acids and its Applications*. Journal of the American Oil Chemists' Society, 2011. **88**(9): p. 1329-1338.
69. Gardner, R.D., et al., *Use of sodium bicarbonate to stimulate triacylglycerol accumulation in the chlorophyte Scenedesmus sp. and the diatom Phaeodactylum tricornutum*. Journal of Applied Phycology, 2012. **24**(5): p. 1311-1320.
70. Gardner, R., et al., *Cellular Cycling, Carbon Utilization, and Photosynthetic Oxygen Production during Bicarbonate-Induced Triacylglycerol Accumulation in a Scenedesmus sp.* Energies, 2013. **6**(11): p. 6060-6076.
71. Gardner, R., et al., *Medium pH and nitrate concentration effects on accumulation of triacylglycerol in two members of the chlorophyta*. Journal of Applied Phycology, 2010. **23**(6): p. 1005-1016.
72. Guckert, J.B. and K.E. Cooksey, *Triglyceride Accumulation and Fatty Acid Profile Changes in Chlorella (Chlorophyta) During High pH-Induced Cell Cycle Inhibition*. Journal of Phycology, 1990. **26**(1): p. 72-79.
73. Sharma, K.K., H. Schuhmann, and P.M. Schenk, *High Lipid Induction in Microalgae for Biodiesel Production*. Energies, 2012. **5**(5): p. 1532-1553.
74. Valenzuela, J., et al., *Nutrient resupplementation arrests bio-oil accumulation in Phaeodactylum tricornutum*. Appl Microbiol Biotechnol, 2013. **97**(15): p. 7049-59.
75. Gardner, R.D., et al., *Comparison of CO₂ and bicarbonate as inorganic carbon sources for triacylglycerol and starch accumulation in Chlamydomonas reinhardtii*. Biotechnol Bioeng, 2013. **110**(1): p. 87-96.
76. Hunt, R.W., et al., *Effect of biochemical stimulants on biomass productivity and metabolite content of the microalga, Chlorella sorokiniana*. Applied biochemistry and biotechnology, 2010. **162**(8): p. 2400-2414.

77. Griffiths, M. and S. Harrison, *Lipid productivity as a key characteristic for choosing algal species for biodiesel production*. Journal of Applied Phycology, 2009. **21**(5): p. 493-507.
78. Borowitzka, M.A., *Algal biotechnology products and processes — matching science and economics*. Journal of Applied Phycology, 1992. **4**(3): p. 267-279.
79. Stumm, W. and J.J. Morgan, *Aquatic chemistry: chemical equilibria and rates in natural waters*. Vol. 126. 2012: John Wiley & Sons.
80. Jones, B.E., et al., *Microbial diversity of soda lakes*. Extremophiles, 1998. **2**(3): p. 191-200.
81. Kilham, J.M.M.P., *Photosynthetic activity of phytoplankton in tropical African soda lakes*. Limnology and Oceanography, 1974. **19**(5): p. 743-755.
82. Richards, A., *Identification and structural characterization of siderophores produced by halophilic and alkaliphilic bacteria*, in *Department of Chemical Engineering*. 2007, Washington State University.
83. Pick, U., Karni, Leah And Avron, Mokdhay *Determination of Ion Content and Ion Fluxes in the Halotolerant Alga Dunaliella salina*. Plant Physiol., 1986. **82**: p. 91-96.
84. Cooksey, K.E., *Regulation of the initial events in microalgal triacylglycerol (TAG) synthesis: hypotheses*. Journal of Applied Phycology, 2014.
85. Cooksey, K.E., *Regulation of the initial events in microalgal triacylglycerol (TAG) synthesis: hypotheses*. Journal of applied phycology, 2015. **27**(4): p. 1385-1387.
86. Madigan, M.T., Martinko, John M., Stahl, David, A. & Clark, David P., *Brock Biology of Microorganisms*. 13 ed. 2012, Boston: Benjamin Cummings.
87. Cooksey, K.E., et al., *Fluorometric determination of the neutral lipid content of microalgal cells using Nile Red*. Journal of Microbiological Methods, 1987. **6**(6): p. 333-345.
88. Davis, R., A. Aden, and P.T. Pienkos, *Techno-economic analysis of autotrophic microalgae for fuel production*. Applied Energy, 2011. **88**(10): p. 3524-3531.
89. Terry, K.L. and L.P. Raymond, *System design for the autotrophic production of microalgae*. Enzyme and Microbial Technology, 1985. **7**(10): p. 474-487.
90. Williams, P.J.I.B. and L.M.L. Laurens, *Microalgae as biodiesel & biomass feedstocks: Review & analysis of the biochemistry, energetics & economics*. Energy & Environmental Science, 2010. **3**(5): p. 554.
91. Amin, S., *Review on biofuel oil and gas production processes from microalgae*. Energy Conversion and Management, 2009. **50**(7): p. 1834-1840.
92. Hise, A.M., et al., *Evaluating the relative impacts of operational and financial factors on the competitiveness of an algal biofuel production facility*. Bioresour Technol, 2016. **220**: p. 271-281.
93. Burns, N.A., *Biomass--The next revolution in surfactants?* Inform, 2010. **21**(727-729): p. 779.
94. Ahmad, A.L., et al., *Microalgae as a sustainable energy source for biodiesel production: A review*. Renewable and Sustainable Energy Reviews, 2011. **15**(1): p. 584-593.
95. King, J.M.D., *Oil's tipping point has passed*. Nature, 2012. **481**: p. 434-435.
96. Hamilton, J.D., *Causes and Consequences of the Oil Shock of 2007–08*. Brookings Papers on Economic Activity, 2009. **Spring**: p. 215-283.

97. Finer, M., et al., *Oil and gas projects in the Western Amazon: threats to wilderness, biodiversity, and indigenous peoples*. PLoS One, 2008. **3**(8): p. e2932.
98. Geltman, E.A.G., *Oil & Gas Drilling in National Parks*. Natural Resources Journal, Winter 2016. **56**(145).
99. Geltman, E.A.G., *Oil & Gas Drilling in National Parks*. Natural Resources Journal, 2016. **56**(1): p. 145-192.
100. Mata, T.M., A.A. Martins, and N.S. Caetano, *Microalgae for biodiesel production and other applications: A review*. Renewable and Sustainable Energy Reviews, 2010. **14**(1): p. 217-232.
101. Moll, K.M., Pedersen, T.C., Gardner, R.D., & Peyton, B.M., *Biodiesel (Microalgae)*, in *Extremophilic Microbial Processing of Lignocellulosic Feedstocks to Biofuels, Value-Added Products, and Usable Power*, D.R. Sani, Editor. 2018, Springer: New York, NY. p. 63-78.
102. Nalley, J.O., M. Stockenreiter, and E. Litchman, *Community Ecology of Algal Biofuels: Complementarity and Trait-Based Approaches*. Industrial Biotechnology, 2014. **10**(3): p. 191-201.
103. Xu, C., et al., *The Use of the Schizonticidal Agent Quinine Sulfate to Prevent Pond Crashes for Algal-Biofuel Production*. Int J Mol Sci, 2015. **16**(11): p. 27450-6.
104. Carney, L.T., et al., *Pond Crash Forensics: Presumptive identification of pond crash agents by next generation sequencing in replicate raceway mass cultures of *Nannochloropsis salina**. Algal Research, 2016. **17**: p. 341-347.
105. Park, S., et al., *The Selective Use of Hypochlorite to Prevent Pond Crashes for Algae-Biofuel Production*. Water Environment Research, 2016. **88**(1): p. 70-78.
106. McBride, R.C., et al., *Contamination Management in Low Cost Open Algae Ponds for Biofuels Production*. Industrial Biotechnology, 2014. **10**(3): p. 221-227.
107. Gardner, R., et al., *Use of sodium bicarbonate to stimulate triacylglycerol accumulation in the chlorophyte *Scenedesmus* sp. and the diatom & *Phaeodactylum tricornutum**. Journal of Applied Phycology, 2012: p. 1-10.
108. Pedersen, T.C., et al., *Assessment of *Nannochloropsis gaditana* growth and lipid accumulation with increased inorganic carbon delivery*. Journal of Applied Phycology, 2018. **30**(4): p. 2155-2166.
109. Doemel, W.N.B., T.D., *The Physiological Ecology of *Cyanothidium caldarium**. Journal of General Microbiology, 1971. **67**: p. 17-32.
110. Skorupa, D.J., et al., *In situ gene expression profiling of the thermoacidophilic alga *Cyanidioschyzon* in relation to visible and ultraviolet irradiance*. Environ Microbiol, 2014. **16**(6): p. 1627-41.
111. Oren, A., *The ecology of *Dunaliella* in high-salt environments*. Journal of Biological Research-Thessaloniki, 2014. **21**(1): p. 23.
112. Palmisano, A.C. and C.W. Sullivan, *Growth, metabolism, and dark survival in sea ice microalgae*, in *Sea ice biota*. 2018, CRC Press. p. 131-146.
113. Wensel, P., et al., *Isolation, characterization, and validation of oleaginous, multi-trophic, and haloalkaline-tolerant microalgae for two-stage cultivation*. Algal Research, 2014. **4**: p. 2-11.

114. Vadlamani, A., et al., *Cultivation of microalgae at extreme alkaline pH conditions: a novel approach for biofuel production*. ACS Sustainable Chemistry & Engineering, 2017. **5**(8): p. 7284-7294.
115. Gardner, R., et al., *Medium pH and nitrate concentration effects on accumulation of triacylglycerol in two members of the chlorophyta*. Journal of Applied Phycology, 2011. **23**(6): p. 1005-1016.
116. Moll, K., *Diatom Biofuels: Optimizing Nutrient Requirements for Growth and Lipid Accumulation in YNP Isolate RGd-1*, in *Microbiology*. 2012, Montana State University: Bozeman, MT. p. 170.
117. Bischoff, H.W. and H.C. Bold, *Some soil algae from enchanted rock and related Algal species*. Publication / The University of Texas;no. 6318. 1963, Austin: [s.n.].
118. Provasoli, L., J.J.A. McLaughlin, and M.R. Droop, *The development of artificial media for marine algae*. Archives of Microbiology, 1957. **25**(4): p. 392-428.
119. Andersen, R.e., *Algal Culturing Techniques*. 2005, San Francisco: Academic Press.
120. Bold, H.C., *The Morphology of Chlamydomonas chlamydogama, Sp. Nov.* Bulletin of the Torrey Botanical Club, Mar-Apr., 1949. **76**(2): p. 101-108.
121. Kuenen, L.A.R.a.J.G., *Aerobic denitrification a controversy revived*. Archives of Microbiology, 1984. **139**: p. 351-354.
122. Bowen De Leon, K., B.D. Ramsay, and M.W. Fields, *Quality-score refinement of SSU rRNA gene pyrosequencing differs across gene region for environmental samples*. Microb Ecol, 2012. **64**(2): p. 499-508.
123. Bell, T.A.S., et al., *Contributions of the microbial community to algal biomass and biofuel productivity in a wastewater treatment lagoon system*. Algal Research, 2019. **39**.
124. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data*. Nature Methods, 2010. **7**(5): p. 335-336.
125. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences*. Bioinformatics, 2006. **22**(13): p. 1658-9.
126. Ybanez, A.P., M. Sashika, and H. Inokuma, *The phylogenetic position of Anaplasma bovis and inferences on the phylogeny of the genus Anaplasma*. Journal of Veterinary Medical Science, 2013: p. 13-0411.
127. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies*. Bioinformatics, 2014. **30**(9): p. 1312-3.
128. White, T.J., Bruns, T., Lee, S., & Taylor, J. , *Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics*, in *PCR Protocols: A Guide to Methods and Applications*, M.A. Innis, Gelfand, D.H., Sninsky, J.J. & White, T.J., Editor. 1990, Academic Press, Inc.: San Diego. p. 315-322.
129. Gardes, M., et al., *Identification of indigenous and introduced symbiotic fungi in ectomycorrhizae by amplification of nuclear and mitochondrial ribosomal DNA*. Canadian Journal of Botany, 1991. **69**(1): p. 180-190.
130. Gardes, M. and T.D. Bruns, *ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts*. Molecular Ecology, 1993. **2**(2): p. 113-118.
131. Mitchell, T.G., et al., *Unique oligonucleotide primers in PCR for identification of Cryptococcus neoformans*. Journal of Clinical Microbiology, 1994. **32**(1): p. 253-255.

132. Chen, W., et al., *A high throughput Nile red method for quantitative measurement of neutral lipids in microalgae*. J Microbiol Methods, 2009. **77**(1): p. 41-7.
133. Gardner, R.D., et al., *Comparison of CO₂ and bicarbonate as inorganic carbon sources for triacylglycerol and starch accumulation in Chlamydomonas reinhardtii*. Biotechnology and Bioengineering, 2012.
134. Weiss, T.L., et al., *Colony organization in the green alga Botryococcus braunii (Race B) is specified by a complex extracellular matrix*. Eukaryotic cell, 2012. **11**(12): p. 1424-1440.
135. Natalia Pismenskaya, E.L., Victor Nikonenko, Abdulla El Attar, Bernard Auclair, Gérald Pourcelly, *Dependence of composition of anion-exchange membranes and their electrical conductivity on concen-V*. Journal of Membrane Science, 2001. **181**: p. 185-197.
136. Cheng, P., et al., *The growth, lipid and hydrocarbon production of Botryococcus braunii with attached cultivation*. Bioresour Technol, 2013. **138**: p. 95-100.
137. Vazquez-Duhalt, R. and H. Greppin, *Growth and production of cell constituents in batch cultures of botryococcus sudeticus*. Phytochemistry, 1987. **26**(4): p. 885-889.
138. Ruangsomboon, S., *Effect of light, nutrient, cultivation time and salinity on lipid production of newly isolated strain of the green microalga, Botryococcus braunii KMITL 2*. Bioresource Technology, 2012. **109**: p. 261-265.
139. Sydney, E.d., et al., *Screening of microalgae with potential for biodiesel production and nutrient removal from treated domestic sewage*. Applied Energy, 2011. **88**(10): p. 3291-3294.
140. Mata, T.M., A.n.A. Martins, and N.S. Caetano, *Microalgae for biodiesel production and other applications: A review*. Renewable and Sustainable Energy Reviews, 2010. **14**(1): p. 217-232.
141. Zhou, W., et al., *Local bioprospecting for high-lipid producing microalgal strains to be grown on concentrated municipal wastewater for biofuel production*. Bioresour Technol, 2011. **102**(13): p. 6909-19.
142. Liu, J., et al., *Aerated swine lagoon wastewater: a promising alternative medium for Botryococcus braunii cultivation in open system*. Bioresour Technol, 2013. **139**: p. 190-4.
143. Ji, M.K., et al., *Effect of mine wastewater on nutrient removal and lipid production by a green microalga Micratinium reisseri from concentrated municipal wastewater*. Bioresour Technol, 2014. **157**: p. 84-90.
144. Eustance, E., et al., *Growth, nitrogen utilization and biodiesel potential for two chlorophytes grown on ammonium, nitrate or urea*. Journal of Applied Phycology, 2013. **25**(6): p. 1663-1677.
145. Nelson, I.W.A.F.T.L.a.Y., *Algae Grown on Dairy and Municipal Wastewater for Simultaneous Nutrient Removal and Lipid Production for Biofuel Feedstock*. Journal of Environmental Engineering, 2009. **135**(11): p. 1115-1122.
146. Ann C. Wilkie, W.W.M., *Recovery of dairy manure nutrients by benthic freshwater algae*. Bioresour Technol, 2002. **84**: p. 81-91.
147. Cheng, D.L., et al., *Microalgae biomass from swine wastewater and its conversion to bioenergy*. Bioresour Technol, 2019. **275**: p. 109-122.
148. Gopalakrishnan, K., J. Roostaei, and Y. Zhang, *Mixed culture of Chlorella sp. and wastewater wild algae for enhanced biomass and lipid accumulation in artificial*

- wastewater medium. *Frontiers of Environmental Science & Engineering*, 2018. **12**(4): p. 14.
149. Kothari, R., et al., *Experimental study for growth potential of unicellular alga Chlorella pyrenoidosa on dairy waste water: an integrated approach for treatment and biofuel production*. *Bioresour Technol*, 2012. **116**: p. 466-70.
 150. Farooq Ahmad, A.U.K.a.A.Y., *Uptake of Nutrients from Municipal Wastewater and Biodiesel Production by Mixed Algae Culture*. *Pakistan Journal of Nutrition*, 2012. **11**(7).
 151. Guarnieri, M.T., et al., *Genome Sequence of the Oleaginous Green Alga, Chlorella vulgaris UTEX 395*. *Front Bioeng Biotechnol*, 2018. **6**: p. 37.
 152. Lohman, E.J., et al., *Optimized inorganic carbon regime for enhanced growth and lipid accumulation in Chlorella vulgaris*. *Biotechnol Biofuels*, 2015. **8**: p. 82.
 153. Bernstein, H.C., et al., *Direct measurement and characterization of active photosynthesis zones inside wastewater remediating and biofuel producing microalgal biofilms*. *Bioresour Technol*, 2014. **156**: p. 206-15.
 154. Smetacek, V., et al., *Deep carbon export from a Southern Ocean iron-fertilized diatom bloom*. *Nature*, 2012. **487**(7407): p. 313-9.
 155. A. Subramaniam, P.L.Y., E. J. Carpenter, C. Mahaffey, K. Björkman, S. Cooley, A. B. Kustka, J. P. Montoya, S. A. Sanudo-Wilhelmy, R. Shipe, and D. G. Capone, *Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean*. *PNAS*, 2008. **105**(30): p. 10460–10465.
 156. Gao, B., et al., *Co-production of lipids, eicosapentaenoic acid, fucoxanthin, and chrysolaminarin by Phaeodactylum tricornutum cultured in a flat-plate photobioreactor under varying nitrogen conditions*. *Journal of Ocean University of China*, 2017. **16**(5): p. 916-924.
 157. Lohman, E.J., et al., *An efficient and scalable extraction and quantification method for algal derived biofuel*. *J Microbiol Methods*, 2013. **94**(3): p. 235-44.
 158. Nurachman, Z., et al., *Oil from the tropical marine benthic-diatom Navicula sp.* *Appl Biochem Biotechnol*, 2012. **168**(5): p. 1065-75.
 159. Seymour, J.R., et al., *Zooming in on the phycosphere: the ecological interface for phytoplankton-bacteria relationships*. *Nat Microbiol*, 2017. **2**: p. 17065.
 160. Johansson, O.N., et al., *Friends With Benefits: Exploring the Phycosphere of the Marine Diatom Skeletonema marinoi*. *Front Microbiol*, 2019. **10**: p. 1828.
 161. Grossart, H.-P., et al., *Marine diatom species harbour distinct bacterial communities*. *Environmental Microbiology*, 2005. **7**(6): p. 860-873.
 162. Droop, M.R., *Vitamins, phytoplankton and bacteria: symbiosis or scavenging?* *Journal of Plankton Research*, 2007. **29**(2): p. 107-113.
 163. Grossart, H.-P., G. Czub, and M. Simon, *Algae-bacteria interactions and their effects on aggregation and organic matter flux in the sea*. *Environmental Microbiology*, 2006. **8**(6): p. 1074-1084.
 164. Pisman, T.I., Y.V. Galayda, and N.S. Loginova, *Population dynamics of an algal–bacterial cenosis in closed ecological system*. *Advances in Space Research*, 2005. **35**(9): p. 1579-1583.
 165. Hutchins, D.A. and K.W. Bruland, *Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime*. *Nature*, 1998. **393**(6685): p. 561-564.

166. Mark Hildebrand, A.K.D., Sarah R Smith, Jesse C Traller & Raffaella Abbriano, *The place of diatoms in the biofuels industry.pdf*. Biofuels, 2012. **3**(2): p. 221-240.
167. Sison-Mangus, M.P., et al., *Host-specific adaptation governs the interaction of the marine diatom, Pseudo-nitzschia and their microbiota*. ISME J, 2014. **8**(1): p. 63-76.
168. William S. and Helene Feil, A.C., *Bacterial genomic DNA isolation using CTAB*, M.H. 11-12-12, Editor. 2004, Joint Genome Institute, Department of Energy.
169. Hunken, M., J. Harder, and G.O. Kirst, *Epiphytic bacteria on the Antarctic ice diatom Amphiprora kufferathii Manguin cleave hydrogen peroxide produced during algal photosynthesis*. Plant Biol (Stuttg), 2008. **10**(4): p. 519-26.
170. Droop, M., *A procedure for routine purification of algal cultures with antibiotics*. British Phycological Bulletin, 1967. **3**(2): p. 295-297.
171. Han, X.Y. and R.A. Andrade, *Brevundimonas diminuta infections and its resistance to fluoroquinolones*. J Antimicrob Chemother, 2005. **55**(6): p. 853-9.
172. Kim, K.E., et al., *Long-read, whole-genome shotgun sequence data for five model organisms*. Sci Data, 2014. **1**: p. 140045.
173. Ye, C., et al., *DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies*. Sci Rep, 2016. **6**: p. 31900.
174. Chin, C.S., et al., *Phased diploid genome assembly with single-molecule real-time sequencing*. Nat Methods, 2016. **13**(12): p. 1050-1054.
175. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. bioRxiv, 2016.
176. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing*. J Comput Biol, 2012. **19**(5): p. 455-77.
177. Huang, X., *CAP3: A DNA Sequence Assembly Program*. Genome Research, 1999. **9**(9): p. 868-877.
178. Simpson, J.T. and R. Durbin, *Efficient de novo assembly of large genomes using compressed data structures*. Genome Res, 2012. **22**(3): p. 549-56.
179. Zdobnov, E.M., et al., *OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs*. Nucleic Acids Res, 2017. **45**(D1): p. D744-D749.
180. Stanke, M., et al., *AUGUSTUS: ab initio prediction of alternative transcripts*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W435-9.
181. Stanke, M., et al., *Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources*. BMC Bioinformatics, 2006. **7**: p. 62.
182. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
183. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects*. BMC Bioinformatics, 2011. **12**: p. 491.
184. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome*. Nat Biotechnol, 2011. **29**(7): p. 644-52.
185. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis*. Nat Protoc, 2013. **8**(8): p. 1494-512.

186. Campbell, M.S., et al., *Genome Annotation and Curation Using MAKER and MAKER-P*. Curr Protoc Bioinformatics, 2014. **48**: p. 4 11 1-4 11 39.
187. Cantarel, B.L., et al., *MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes*. Genome Research, 2008. **18**(1): p. 188-196.
188. Korf, I., *Gene finding in novel genomes*. BMC Bioinformatics, 2004. **5**(1): p. 59.
189. Humann, J.L., et al., *Structural and Functional Annotation of Eukaryotic Genomes with GenSAS*, in *Gene Prediction: Methods and Protocols*, M. Kollmar, Editor. 2019, Springer New York: New York, NY. p. 29-51.
190. Kanehisa, M. and Y. Sato, *KEGG Mapper for inferring cellular functions from protein sequences*. Protein Science, 2020. **29**(1): p. 28-35.
191. Perez, N., M. Gutierrez, and N. Vera, *Computational Performance Assessment of k-mer Counting Algorithms*. J Comput Biol, 2016. **23**(4): p. 248-55.
192. Wu, Y.W., *ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes*. BMC Genomics, 2018. **19**(Suppl 1): p. 921.
193. El-Gebali, S., et al., *The Pfam protein families database in 2019*. Nucleic Acids Res, 2019. **47**(D1): p. D427-D432.
194. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability*. Mol Biol Evol, 2013. **30**(4): p. 772-80.
195. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses*. Bioinformatics, 2009. **25**(15): p. 1972-3.
196. Rambaut, A., *FigTree v1. 4.0. A graphical viewer of phylogenetic trees*. See <http://tree.bio.ed.ac.uk/software/figtree/>, 2012.
197. Bell, T.A., et al., *A Lipid-Accumulating Alga Maintains Growth in Outdoor, Alkaliphilic Raceway Pond with Mixed Microbial Communities*. Front Microbiol, 2015. **6**: p. 1480.
198. McKay, L.J., et al., *Occurrence and expression of novel methyl-coenzyme M reductase gene (mcrA) variants in hot spring sediments*. Sci Rep, 2017. **7**(1): p. 7252.
199. Ounit, R., et al., *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers*. BMC Genomics, 2015. **16**: p. 236.
200. Cole, J.R., et al., *Ribosomal Database Project: data and tools for high throughput rRNA analysis*. Nucleic Acids Res, 2014. **42**(Database issue): p. D633-42.
201. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements*. Nat Methods, 2015. **12**(4): p. 357-60.
202. Robert Henschel, P.M.N., Matthias Lieber, Brian J. Haas, Le-Shin Wu & Richard D. LeDuc, *Trinity RNA-Seq Assembler Performance Optimization.pdf*, in *XSEDE12*. July 16 - 20 2012: Chicago, Illinois, USA.
203. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report*. Bioinformatics, 2016. **32**(19): p. 3047-8.
204. Smith, S.R., R.M. Abbriano, and M. Hildebrand, *Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways*. Algal Research, 2012. **1**(1): p. 2-16.
205. Wei Zhao, X.H., Katherine A Hoadley, Joel S Parker, David Neil Hayes & Charles M Perou, *Comparison of RNASeq by polyA capture ribosomal RNA depletion and DNA microarray for expression profiling*. BMC Genomics, 2014. **15**(419): p. 1-11.

206. Hrdlickova, R., M. Toloue, and B. Tian, *RNA-Seq methods for transcriptome analysis*. Wiley Interdisciplinary Reviews: RNA, 2017. **8**(1): p. e1364.
207. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
208. Simpson, J.T., *Exploring genome characteristics and sequence quality without a reference*. Bioinformatics, 2014. **30**(9): p. 1228-1235.
209. Armbrust, E.V., *The life of diatoms in the world's oceans*. Nature, 2009. **459**(7244): p. 185-92.
210. Delcher, A.L., S.L. Salzberg, and A.M. Phillippy, *Using MUMmer to identify similar regions in large sequence sets*. Current protocols in bioinformatics, 2003(1): p. 10.3. 1-10.3. 18.
211. White, D., *The Physiology and Biochemistry of Prokaryotes*. Third Edition ed. 2007, New York: Oxford University Press.
212. Grisham, R.H.G.C.M., *Biochemistry*. Fourth Edition ed. 2010, University of Virginia: Brooks/Cole.
213. Kajikawa, M., et al., *Production of ricinoleic acid-containing monoester triacylglycerides in an oleaginous diatom, Chaetoceros gracilis*. Scientific Reports, 2016. **6**(1).
214. Sally A. Ong, T.P.a.J.B.N., *Agrobactin, a Siderophore from Agrobacterium tumefaciens*. The Journal of Biological Chemistry, 1979. **254**(6): p. 1860-1865.
215. Mehansho, H., *Iron Fortification Technology Development: New Approaches*. The Journal of Nutrition, 2006. **136**(4): p. 1059-1063.
216. BRITT A. HOLME N, J.D.S., DOUGLAS C. NELSON, and WILLIAM H. CASEY, *Hydroxamate siderophores, cell growth and Fe(III) cycling in two anaerobic iron oxide media containing Geobacter metallireducens*. Geochimica et Cosmochimica Acta, 1999. **63**(2): p. 227-239.
217. Penyalver, R., et al., *Iron-binding compounds from Agrobacterium spp.: biological control strain Agrobacterium rhizogenes K84 produces a hydroxamate siderophore*. Appl Environ Microbiol, 2001. **67**(2): p. 654-64.
218. Park, Y., et al., *Growth promotion of Chlorella ellipsoidea by co-inoculation with Brevundimonas sp. isolated from the microalga*. Hydrobiologia, 2007. **598**(1): p. 219-228.
219. Schubbe, S., et al., *Complete genome sequence of the chemolithoautotrophic marine magnetotactic coccus strain MC-1*. Appl Environ Microbiol, 2009. **75**(14): p. 4835-52.
220. Madhaiyan, M., et al., *Arachidicoccus rhizosphaerae gen. nov., sp. nov., a plant-growth-promoting bacterium in the family Chitinophagaceae isolated from rhizosphere soil*. Int J Syst Evol Microbiol, 2015. **65**(Pt 2): p. 578-86.
221. Tu, J., et al., *The siderophore-interacting protein is involved in iron acquisition and virulence of Riemerella anatipestifer strain CH3*. Veterinary Microbiology, 2014. **168**(2): p. 395-402.
222. Karen M. Moll, N.D., Thiruvarangan Ramaraj, Joann Mudge, Brent M. Peyton *GENOME SEQUENCE FOR AN EXTREMOPHILIC BREVUNDIMONAS STRAIN*. Microbiology Resource Announcements, In Progress.

223. Ghosh, P., et al., *Bacterial ability in AsIII oxidation and AsV reduction: Relation to arsenic tolerance, P uptake, and siderophore production*. Chemosphere, 2015. **138**: p. 995-1000.
224. Valenzuela, J., Mazurie, A., Carlson, R.P., Gerlach, R., Cooksey, K.E., Peyton, B.M & Fields, M.W., *Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum*. Biotechnology for biofuels, 2012. **5**(40): p. 10.1186/1754-6834-5-40.
225. Erb, T.J., et al., *Synthesis of C5-dicarboxylic acids from C2-units involving crotonyl-CoA carboxylase/reductase: the ethylmalonyl-CoA pathway*. Proc Natl Acad Sci U S A, 2007. **104**(25): p. 10631-6.
226. Seyedsayamdost, M.R., et al., *The Jekyll-and-Hyde chemistry of Phaeobacter gallaeciensis*. Nat Chem, 2011. **3**(4): p. 331-5.
227. Chen, G., L. Zhao, and Y. Qi, *Enhancing the productivity of microalgae cultivated in wastewater toward biofuel production: a critical review*. Applied Energy, 2015. **137**: p. 282-291.
228. Singh, N., et al., *Brevundimonas diminuta mediated alleviation of arsenic toxicity and plant growth promotion in Oryza sativa L*. Ecotoxicol Environ Saf, 2016. **125**: p. 25-34.
229. Schwyn, B. and J.B. Neilands, *Universal chemical assay for the detection and determination of siderophores*. Analytical Biochemistry, 1987. **160**(1): p. 47-56.
230. Tan, Y., et al., *The Low Conductivity of Geobacter uraniireducens Pili Suggests a Diversity of Extracellular Electron Transfer Mechanisms in the Genus Geobacter*. Front Microbiol, 2016. **7**: p. 980.
231. Anderson, R.T., et al., *Stimulating the in situ activity of Geobacter species to remove uranium from the groundwater of a uranium-contaminated aquifer*. Appl Environ Microbiol, 2003. **69**(10): p. 5884-91.
232. Kroth, P.G., et al., *A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis*. PLoS One, 2008. **3**(1): p. e1426.
233. Davis, A., et al., *Clarification of Photorespiratory Processes and the Role of Malic Enzyme in Diatoms*. Protist, 2017. **168**(1): p. 134-153.
234. Wang, M., et al., *PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations*. BMC genomics, 2015. **16**(1): p. 214.
235. Seemann, T., *Prokka: rapid prokaryotic genome annotation*. Bioinformatics, 2014. **30**(14): p. 2068-9.
236. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)*. Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.
237. Brettin, T., et al., *RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes*. Sci Rep, 2015. **5**: p. 8365.
238. Soria-Dengg, S.a.H., U., *Ferrioxamines B and E as iron sources for the marine diatom Phaeodactylum tricornutum*. Marine Ecology Progress Series, 1995. **127**: p. 269-277.

239. Soria-Dengg, S., Reissbrodt, R. & Horstmann, U., *Siderophores in marine coastal waters and their relevance for iron uptake by phytoplankton experiments with the diatom Phaeodactylum tricornutum.pdf*. Marine Ecology - Progress Series, 2001. **220**(73-82).
240. Gurney, K.R., et al., *High Resolution Fossil Fuel Combustion CO₂ Emission Fluxes for the United States*. Environmental Science & Technology, 2009. **43**(14): p. 5535-5541.
241. Hutchins, D.A., et al., *Climate change microbiology - problems and perspectives*. Nat Rev Microbiol, 2019. **17**(6): p. 391-396.
242. Kuma, K., et al., *Effect of Hydroxamate Ferrisiderophore Complex (Ferrichrome) on Iron Uptake and Growth of a Coastal Marine Diatom, Chaetoceros sociale*. Limnology and Oceanography, 2000. **45**(6): p. 1235-1244.
243. Kustka, A.B., A.E. Allen, and F.M.M. Morel, *Sequence analysis and transcriptional regulation of iron acquisition genes in two marine diatoms*. Journal of Phycology, 2007. **43**(4): p. 715-729.
244. Elena Kazamia, R.S., Javier Paz-Yepes, Richard G. Dorrell, and J.M. Fabio Rocha Jimenez Vieira, Joe Morrissey, Sébastien Leon, France Lam, Eric Pelletier, Jean-Michel Camadro, Chris Bowler, Emmanuel Lesuisse, *Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms*. Science Advances, 2018. **4**(5): p. 1-14.
245. Siebers, M., et al., *Lipids in plant-microbe interactions*. Biochim Biophys Acta, 2016. **1861**(9 Pt B): p. 1379-1395.
246. Hwangbo, M. and K.H. Chu, *Recent advances in production and extraction of bacterial lipids for biofuel production*. Sci Total Environ, 2020. **734**: p. 139420.
247. De Riso, V., et al., *Gene silencing in the marine diatom Phaeodactylum tricornutum*. Nucleic Acids Res, 2009. **37**(14): p. e96.
248. Chisti, Y., *Constraints to commercialization of algal fuels*. J Biotechnol, 2013. **167**(3): p. 201-14.
249. Huysman, M.J., et al., *AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum)*. Plant Cell, 2013. **25**(1): p. 215-28.
250. Huysman, M.J., W. Vyverman, and L. De Veylder, *Molecular regulation of the diatom cell cycle*. J Exp Bot, 2014. **65**(10): p. 2573-84.
251. Coesel, S., et al., *Diatom PtCPF1 is a new cryptochrome/photolyase family member with DNA repair and transcription regulation activity*. EMBO reports, 2009. **10**(6): p. 655-661.
252. Anne Jungandreas1, Benjamin Schellenberger Costa, Torsten Jakob, Martin von Bergen and C.W. Sven Baumann, *The Acclimation of Phaeodactylum tricornutum to Blue and Red Light Does Not Influence the Photosynthetic Light Reaction but Strongly Disturbs the Carbon Allocation Pattern*. PLOS One, 2014. **9**(8): p. 1-14.
253. Benjamin Schellenberger Costa, M.S., Anne Jungandreas, Carolina Rio Bartulos, Ansgar Gruber, Torsten Jakob, Peter G. Kroth , Christian Wilhelm, *Aureochrome 1a is involved in the Photoacclimation of the Diatom Phaeodactylum tricornutum.pdf*. PLoS One, 2013. **8**(9): p. 1-14.
254. Beel, B., et al., *A flavin binding cryptochrome photoreceptor responds to both blue and red light in Chlamydomonas reinhardtii*. The Plant Cell Online, 2012. **24**(7): p. 2992-3008.

255. Armbrust, E.V., *The life of diatoms in the world's oceans*. Nature, 2009. **459**(7244): p. 185-192.
256. Cao, S., J. Wang, and D. Chen, *Settlement and cell division of diatom Navicula can be influenced by light of various qualities and intensities*. J Basic Microbiol, 2013. **53**(11): p. 884-94.
257. Xuhong Yu, H.L., John Klejnot, and Chentao Lin, *The Cryptochrome Blue Light Receptors*, in *The Arabidopsis Book*. 2010. p. 1-27.
258. Xuhong Yu, H.L., John Klejnot, and Chentao Lin, *The Cryptochrome Blue Light Receptors.pdf*. The Arabidopsis book/American Society of Plant Biologists, 2010: p. 2-27.
259. Su, Y., N. Lundholm, and M. Ellegaard, *The effect of different light regimes on diatom frustule silicon concentration*. Algal Research, 2018. **29**: p. 36-40.
260. Su, Y., Nina Lundholm, Søren M. M. Friis & Marianne Ellegaard, *Implications for photonic applications of diatom growth and frustule nanostructure changes in response to different light wavelengths*. Nano Research, 2015: p. 1-10.
261. Huysman, M., et al., *Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling*. Genome Biology, 2010. **11**(2): p. R17.
262. Kafri, R., et al., *Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle*. Nature, 2013. **494**(7438): p. 480-3.
263. Ritchie, R.J., *Consistent sets of spectrophotometric chlorophyll equations for acetone, methanol and ethanol solvents*. Photosynth Res, 2006. **89**(1): p. 27-41.
264. Ritchie, R., *Universal chlorophyll equations for estimating chlorophylls a, b, c, and d and total chlorophylls in natural assemblages of photosynthetic organisms using acetone, methanol, or ethanol solvents*. Photosynthetica, 2008. **46**(1): p. 115-126.
265. Beale, G.L.M.a.S., *Blue light regulated expression of genes for two early steps of chlorophyll biosynthesis in chlamydomonas reinhardtii*. Plant Physiol, 1995. **109**: p. 471-479.
266. Blankenship, R.E., *Molecular Mechanisms of Photosynthesis*. 2002, Ames, Iowa, USA: Iowa State University Press A Blackwell Science Company.
267. H. K. Lichtenthaler, G.K., U. Prenzel, C. Buschmann, and D. Meier, *Adaptation of chloroplast ultrastructure and of chlorophyll protein levels to high light and low light growth conditions*. Zeitschrift für Naturforschung, 1982. **37**: p. 464-475.
268. Needoba, J.A. and P.J. Harrison, *INFLUENCE OF LOW LIGHT AND A LIGHT: DARK CYCLE ON NO₃- UPTAKE, INTRACELLULAR NO₃-, AND NITROGEN ISOTOPE FRACTIONATION BY MARINE PHYTOPLANKTONI*. Journal of Phycology, 2004. **40**(3): p. 505-516.
269. Vadiveloo, A., et al., *Effect of different light spectra on the growth and productivity of acclimated Nannochloropsis sp. (Eustigmatophyceae)*. Algal Research, 2015. **8**: p. 121-127.
270. Kenneth Eskins, C.Z.J.a.R.S., *Light-quality and irradiance effects on pigments, light-harvesting proteins and Rubisco activity in a chlorophyll- and light- harvesting-deficient soybean mutant*. Physiologia Plantarum, 1991. **83**: p. 47-53.
271. Marc Valls, V. and c.d. Lorenzo, *Exploiting the genetic and biochemical capacities of bacteria for the remediation of heavy metal pollution*. FEMS Microbiology Reviews, 2002. **26**: p. 327-338.

272. Anindita Mitra, S.C., and Dharmendra K. Gupta, *Uptake, Transport, and Remediation of Arsenic by Algae and Higher Plants*, in *Arsenic Contamination in the Environment*, S.C. D.K. Gupta, Editor. 2017, Springer International Publishing. p. 145-169.
273. *Arsenic Contamination in the Environment*. 2017, UK: Springer.
274. Tripathi, R.D., et al., *Arsenic hazards: strategies for tolerance and remediation by plants*. Trends Biotechnol, 2007. **25**(4): p. 158-65.
275. Qin, J., et al., *Biotransformation of arsenic by a Yellowstone thermoacidophilic eukaryotic alga*. Proc Natl Acad Sci U S A, 2009. **106**(13): p. 5213-7.
276. Burke, D.L.J.a.R.M., *BIOLOGICAL MEDIATION OF CHEMICAL SPECIATION II. ARSENATE REDUCTION DURING MARINE PHYTOPLANKTON BLOOMS*. Chemosphere, 1978. **8**: p. 645-648.
277. Sanders, J.G., *Effects of arsenic speciation*. J. Phycol., 1979. **15**: p. 424-484.
278. Toshikazu Kaise, M.O., Takao Nozaki, Kazuhisa Saitoh, Teruaki Sakurai, Chiyo Matsubara, Chuichi Watanabe and Ken'ichi Hanaoka, *Biomethylation of Arsenic in an Arsenic-rich Freshwater Environment*. Applied Organometallic Chemistry, 1997. **11**: p. 297-304.
279. Jie Qin, C.R.L., Chungang Yuan, X. Chris Le, Timothy R. McDermott, and Barry P. Rosen, *Biotransformation of arsenic by a Yellowstone thermoacidophilic eukaryotic alga*. PNAS, 2009. **106**(13): p. 5213-5217.
280. Wang, S. and C.N. Mulligan, *Occurrence of arsenic contamination in Canada: sources, behavior and distribution*. Sci Total Environ, 2006. **366**(2-3): p. 701-21.
281. A. G. Howard, S.D.W.C.D.K.E.E.A., and D. A. Purdie, *Arsenic Speciation and Seasonal Changes in Nutrient Availability and Micro-plankton Abundance in Southampton Water, U.K.* Estuarine Coastal and Shelf Science, 1995. **40**: p. 435-450.
282. Andreae, M.O., *Distribution and speciation of arsenic in natural waters and some marine algae*. Deep Sea Research, 1978. **25**(4): p. 391-402.
283. Meharg, A.A. and A. Raab, *Getting to the bottom of arsenic standards and guidelines*. Environ Sci Technol, 2010. **44**(12): p. 4395-9.
284. *National Primary Drinking Water Regulations; Arsenic and Clarifications to Compliance and New Source Contaminants Monitoring; Final Rule*. January 22, 2001. p. 6976-7066.
285. Barral-Fraga, L., et al., *Short-term arsenic exposure reduces diatom cell size in biofilm communities*. Environ Sci Pollut Res Int, 2016. **23**(5): p. 4257-70.
286. Cibik, J.G.S.S.J., *Adaptive behavior of euryhaline phytoplankton communities to arsenic stress*. Mar. Ecol. Prog. Ser., 1985. **22**: p. 199-205.
287. Healey, D.P.F.P., *EFFECTS OF ARSENATE ON GROWTH AND PHOSPHORUS METABOLISM OF PHYTOPLANKTON*. J Phycol, 1978. **14**: p. 337-341.
288. Oremland, J.R.L.a.R.S., *Microbial Transformations of Arsenic in the Environment From Soda Lakes to Aquifers*. Elements, 2006. **2**: p. 85-90.
289. Sele, V., et al., *Arsenolipids in marine oils and fats: A review of occurrence, chemistry and future research needs*. Food Chemistry, 2012. **133**(3): p. 618-630.
290. Zhu, C. and Y. Lee, *Determination of biomass dry weight of marine microalgae*. Journal of Applied Phycology, 1997. **9**(2): p. 189-194.
291. Chelf, P., *Environmental control of lipid and biomass production in two diatom species*. Journal of Applied Phycology, 1990. **2**(2): p. 121-129.

292. Slaughter, D.C., R.E. Macur, and W.P. Inskeep, *Inhibition of microbial arsenate reduction by phosphate*. Microbiol Res, 2012. **167**(3): p. 151-6.
293. Zienkiewicz, K., et al., *Stress-induced neutral lipid biosynthesis in microalgae - Molecular, cellular and physiological insights*. Biochim Biophys Acta, 2016. **1861**(9 Pt B): p. 1269-1281.
294. Bligh, E.G. and W.J. Dyer, *A rapid method of total lipid extraction and purification*. Canadian journal of biochemistry and physiology, 1959. **37**(8): p. 911-917.
295. Griffiths, M., R. van Hille, and S. Harrison, *Selection of Direct Transesterification as the Preferred Method for Assay of Fatty Acid Content of Microalgae*. Lipids, 2010. **45**(11): p. 1053-1060.
296. Branco, R., A.P. Chung, and P.V. Morais, *Sequencing and expression of two arsenic resistance operons with different functions in the highly arsenic-resistant strain Ochrobactrum tritici SCII24T*. BMC Microbiol, 2008. **8**: p. 95.
297. Zhao, C., et al., *Insights into arsenic multi-operons expression and resistance mechanisms in Rhodopseudomonas palustris CGA009*. Front Microbiol, 2015. **6**: p. 986.
298. Rothstein, A., *Interactions of arsenate with the phosphate-transporting system of yeast*. J Gen Physiol, 1963. **46**: p. 1075-85.
299. SCARBOROUGH, G.A., *The mechanism of arsenate inhibition of the glucose active transport system in Neurospora crassa*. Archives of Biochemistry and Biophysics, 1975. **166**: p. 245-250.
300. Shendure, J. and E.L. Aiden, *The expanding scope of DNA sequencing*. Nature biotechnology, 2012. **30**(11): p. 1084.
301. Mostovoy, Y., et al., *A hybrid approach for de novo human genome sequence assembly and phasing*. Nature methods, 2016. **13**(7): p. 587-590.
302. VanBuren, R., et al., *Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum*. Nature, 2015. **527**(7579): p. 508-511.
303. Bickhart, D.M., et al., *Single-molecule sequencing and conformational capture enable de novo mammalian reference genomes*. bioRxiv, 2016: p. 064352.
304. Ashrafi, H. *Using spinach to compare technologies for whole genome assemblies*. in *Plant and Animal Genome Conference XXIII*. 2015. San Diego, CA.
305. Bertoli, D.J., et al., *The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut*. Nat Genet, 2016. **48**(4): p. 438-46.
306. Chaney, L., et al., *Genome Mapping in Plant Comparative Genomics*. Trends Plant Sci, 2016.
307. Schwartz, D.C., et al., *Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping*. Science, 1993. **262**(5130): p. 110-114.
308. Imelfort, M. and D. Edwards, *De novo sequencing of plant genomes using second-generation technologies*. Brief Bioinform, 2009. **10**(6): p. 609-18.
309. Somes K. Das, M.D.A., Matthew C. Akana, Paru Deshpande, Han Cao and Ming Xiao, *Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes*. Nucleic Acids Res, 2010. **38**(18): p. 1-8.
310. Shelton, J.M., et al., *Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool*. BMC Genomics, 2015. **16**(1): p. 734.
311. Belton, J.M., et al., *Hi-C: a comprehensive technique to capture the conformation of genomes*. Methods, 2012. **58**(3): p. 268-76.

312. Schatz, M.C., J. Witkowski, and W.R. McCombie, *Current challenges in de novo plant genome sequencing and assembly*. Genome biology, 2012. **13**(4): p. 1.
313. Jiao, W.B., et al., *Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data*. Genome Res, 2017.
314. Jiao, Y., et al., *Improved maize reference genome with single-molecule technologies*. Nature, 2017. **546**(7659): p. 524-527.
315. Chin, C.-S., et al., *Phased diploid genome assembly with single-molecule real-time sequencing*. Nat Meth, 2016. **13**(12): p. 1050-1054.
316. Jarvis, D.E., et al., *The genome of Chenopodium quinoa*. Nature, 2017. **542**(7641): p. 307.
317. Zapata, L., et al., *Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms*. Proceedings of the National Academy of Sciences, 2016. **113**(28): p. E4052-E4060.
318. Berlin, K., et al., *Assembling large genomes with single-molecule sequencing and locality-sensitive hashing*. Nature biotechnology, 2015. **33**(6): p. 623-630.
319. Zhang, J., et al., *Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63*. Proceedings of the National Academy of Sciences, 2016. **113**(35): p. E5163-E5171.
320. Daccord, N., et al., *High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development*. Nature genetics, 2017. **49**(7): p. 1099-1106.
321. Du, H., et al., *Sequencing and de novo assembly of a near complete indica rice genome*. Nature communications, 2017. **8**(1): p. 1-12.
322. Reyes-Chin-Wo, S., et al., *Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce*. Nature Communications, 2017. **8**(1): p. 1-11.
323. Bredeson, J.V., et al., *Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity*. Nature biotechnology, 2016. **34**(5): p. 562-570.
324. Pootakham, W., et al., *De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in Hevea species*. Scientific reports, 2017. **7**: p. 41457.
325. Sato, S., et al., *Genome structure of the legume, Lotus japonicus*. DNA research, 2008. **15**(4): p. 227-239.
326. Schmutz, J., et al., *Genome sequence of the palaeopolyploid soybean*. Nature, 2010. **463**(7278): p. 178-83.
327. Young, N.D., et al., *The Medicago genome provides insight into the evolution of rhizobial symbioses*. Nature, 2011. **480**(7378): p. 520-4.
328. Varshney, R.K., et al., *Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement*. Nat Biotechnol, 2013. **31**(3): p. 240-6.
329. Kang, Y.J., et al., *Genome sequence of mungbean and insights into evolution within Vigna species*. Nat Commun, 2014. **5**: p. 5443.
330. Chen, X., et al., *Draft genome of the peanut A-genome progenitor (Arachis duranensis) provides insights into geocarpy, oil biosynthesis, and allergens*. Proc Natl Acad Sci U S A, 2016. **113**(24): p. 6785-90.
331. Li, Y.H., et al., *De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits*. Nat Biotechnol, 2014. **32**(10): p. 1045-52.

332. Gan, X., et al., *Multiple reference genomes and transcriptomes for Arabidopsis thaliana*. Nature, 2011. **477**(7365): p. 419-23.
333. Schatz, M.C., et al., *Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica*. Genome biology, 2014. **15**(11): p. 1.
334. Zhou, P., et al., *Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes*. BMC Genomics, 2017. **18**(1): p. 261.
335. Golicz, A.A., et al., *The pangenome of an agronomically important crop plant Brassica oleracea*. Nature communications, 2016. **7**(1): p. 1-8.
336. Tadege, M., et al., *Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume Medicago truncatula*. Plant J, 2008. **54**(2): p. 335-47.
337. Branca, A., et al., *Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula*. Proceedings of the National Academy of Sciences, 2011. **108**(42): p. E864-E870.
338. Tadege, M., P. Ratet, and K.S. Mysore, *Insertional mutagenesis: a Swiss Army knife for functional genomics of Medicago truncatula*. Trends Plant Sci, 2005. **10**(5): p. 229-35.
339. Tang, H., et al., *An improved genome release (version Mt4. 0) for the model legume Medicago truncatula*. BMC genomics, 2014. **15**(1): p. 1.
340. Cannon, S.B., et al., *Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes*. Proceedings of the National Academy of Sciences, 2006. **103**(40): p. 14959-14964.
341. Blanc, G. and K.H. Wolfe, *Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes*. Plant Cell, 2004. **16**(7): p. 1667-78.
342. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data*. Proceedings of the National Academy of Sciences, 2011. **108**(4): p. 1513-1518.
343. Kamphuis, L.G., et al., *The Medicago truncatula reference accession A17 has an aberrant chromosomal configuration*. New Phytol, 2007. **174**(2): p. 299-303.
344. Chin, C.-S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data*. Nature methods, 2013. **10**(6): p. 563-569.
345. Koren, S., et al., *Hybrid error correction and de novo assembly of single-molecule sequencing reads*. Nature biotechnology, 2012. **30**(7): p. 693-700.
346. Koren, S. and A.M. Phillippy, *One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly*. Current opinion in microbiology, 2015. **23**: p. 110-120.
347. Ribeiro, F.J., et al., *Finished bacterial genomes from shotgun sequence data*. Genome research, 2012. **22**(11): p. 2270-2277.
348. English, A.C., et al., *Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology*. PloS one, 2012. **7**(11): p. e47768.
349. Hongzhi Cao, A.R.H., Dandan Cao, Ernest T Lam, Yuhui Sun, Haodong Huang, Xiao Liu, Liya Lin, Warren Andrew, Saki Chan, Shujia Huang, Xin Tong, Michael Requa, Thomas Anantharaman, Anders Krogh, Huanming Yang, Han Cao and Xun Xu, *Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology*. GigaScience, 2014. **3**(34): p. 1-11.

350. Stankova, H., et al., *BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes*. Plant Biotechnol J, 2016.
351. Meyer, C.A. and X.S. Liu, *Identifying and mitigating bias in next-generation sequencing methods for chromatin biology*. Nat Rev Genet, 2014. **15**(11): p. 709-21.
352. Yaffe, E. and A. Tanay, *Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture*. Nat Genet, 2011. **43**(11): p. 1059-65.
353. Santoferrara, L.F., et al., *De novo transcriptomes of a mixotrophic and a heterotrophic ciliate from marine plankton*. PLoS One, 2014. **9**(7): p. e101418.
354. Simpson, J.T., et al., *ABYSS: a parallel assembler for short read sequence data*. Genome Res, 2009. **19**(6): p. 1117-23.
355. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program*. Genome Research, 1999. **9**(9): p. 868-877.
356. ChristianIseli, C.V.J.P.B., *ESTScan a program for detecting, evaluating, and reconstructing potential coding regions in ESTsequences*. SMB-99 Proceedings, 1999: p. 138-148.
357. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences*. Bioinformatics, 2005. **21**(9): p. 1859-75.
358. Chaisson, M.J.a.T., Glenn, *Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR) application and theory*. BMC Bioinformatics, 2012. **13**(328): p. 1-17.
359. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome*. Nature biotechnology, 2011. **29**(7): p. 644-652.
360. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel*. Bioinformatics, 2003. **19**(Suppl 2): p. ii215-ii225.
361. Kent, W.J., *BLAT—the BLAST-like alignment tool*. Genome research, 2002. **12**(4): p. 656-664.
362. Morgulis, A., et al., *A fast and symmetric DUST implementation to mask low-complexity DNA sequences*. Journal of Computational Biology, 2006. **13**(5): p. 1028-1040.
363. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences*. Nucleic acids research, 1999. **27**(2): p. 573.
364. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future*. Nucleic acids research, 2016. **44**(D1): p. D279-D285.
365. J., S., *A look back at the US Department of Energy's aquatic species program 1988((NREL/TP-580-24190))*.
366. Chisti, Y., *Biodiesel from microalgae*. Biotechnol Adv, 2007. **25**.
367. Georgianna, D.R. and S.P. Mayfield, *Exploiting diversity and synthetic biology for the production of algal biofuels*. Nature, 2012. **488**.
368. Courchesne, N.M.D., et al., *Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches*. J Bacteriol, 2009. **141**.
369. Greenwell, H.C., et al., *Placing microalgae on the biofuels priority list: a review of the technological challenges*. J R Soc Interface, 2010. **7**.
370. Razghafard, R., *Algal biofuels*. Photosynth Res, 2013. **117**.

371. Dismukes, G.C., et al., *Aquatic phototrophs: efficient alternatives to land-based crops for biofuels*. Curr Opin Biotechnol, 2008. **19**.
372. Hu, Q., et al., *Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances*. Plant J, 2008. **54**.
373. Botte, P., et al., *Combined exploitation of CO₂ and nutrient replenishment for increasing biomass and lipid productivity of the marine diatoms *Thalassiosira weissflogii* and *Cyclotella cryptica**. Journal of Applied Phycology, 2017.
374. Mus, F., et al., *Physiological and molecular analysis of carbon source supplementation and pH stress-induced lipid accumulation in the marine diatom *Phaeodactylum tricornutum**. Applied Microbiology and Biotechnology, 2013.
375. Chu, F., et al., *Phosphorus plays an important role in enhancing biodiesel productivity of *Chlorella vulgaris* under nitrogen deficiency*. Bioresour Technol, 2013. **134**.
376. Schnurr, P.J., G.S. Espie, and D.G. Allen, *Algae biofilm growth and the potential to stimulate lipid accumulation through nutrient starvation*. Bioresour Technol, 2013. **136**.
377. Atta, M., et al., *Intensity of blue LED light: a potential stimulus for biomass and lipid content in fresh water microalgae *Chlorella vulgaris**. Bioresour Technol, 2013. **148**.
378. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: the unseen majority*. Proc Natl Acad Sci U S A, 1998. **95**.
379. Pate, R., G. Klise, and B. Wu, *Resource demand implications for US algae biofuels production scale-up*. Appl Energy, 2011. **88**.
380. Aslan, S. and I.K. Kapdan, *Batch kinetics of nitrogen and phosphorus removal from synthetic wastewater by algae*. Ecol Eng, 2006. **28**.
381. Hoffmann, J.P., *Wastewater treatment with suspended and nonsuspended algae*. J Phycol, 2002. **34**.
382. Mallick, N., *Biotechnological potential of immobilized algae for wastewater N, P, and metal removal: a review*. BioMetals, 2002. **15**.
383. Pittman, J.K., A.P. Dean, and O. Osundeko, *The potential of sustainable algal biofuel production using wastewater resources*. Bioresour Technol, 2011. **102**.
384. Gordon, J.M. and J.E.W. Polle, *Ultrahigh bioproductivity from algae*. Appl Microbiol Biotechnol, 2007. **76**.
385. Schuhmann, H., D.K.Y. Lim, and P.M. Schenk, *Perspectives on metabolic engineering for increased lipid contents in microalgae*. Biofuels, 2012. **3**.
386. Valenzuela, J., et al., *Potential role of multiple carbon fixation pathways during lipid accumulation in *Phaeodactylum tricornutum**. Biotechnol Biofuels, 2012. **5**.
387. Odum, E.P., *Trends expected in stressed ecosystems*. Bioscience, 1985. **35**.
388. Ratha, S.K., et al., *Exploring nutritional modes of cultivation for enhancing lipid accumulation in microalgae*. J Basic Microbiol, 2013. **53**.
389. Khozin-Goldberg, I. and Z. Cohen, *The effect of phosphate starvation on the lipid and fatty acid composition of the fresh water eustigmatophyte *Monodus subterraneus**. Phytochemistry, 2006. **67**.
390. Aury, J.M., et al., *Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia**. Nature, 2006. **444**(7116): p. 171-8.
391. Shekh, A.Y., et al., *Stress-induced lipids are unsuitable as a direct biodiesel feedstock: a case study with *Chlorella pyrenoidosa**. Bioresour Technol, 2013. **138**.

392. Msanne, J., et al., *Metabolic and gene expression changes triggered by nitrogen deprivation in the photoautotrophically grown microalgae Chlamydomonas reinhardtii and Coccomyxa sp. C-169*. Phytochemistry, 2012. **75**.
393. Halsey, K.H., et al., *A common partitioning strategy for photosynthetic products in evolutionary distinct phytoplankton species*. New Phytol, 2013. **198**.
394. Brown, A.P., A.R. Slabas, and J.B. Rafferty, *Fatty acid biosynthesis in plants—metabolic pathways, structure and organization*, in *Lipids in photosynthesis*. 2009, Springer. p. 11-34.
395. Wang, Z.T., et al., *Algal lipid bodies: stress induction, purification, and biochemical characterization in wild-type and starchless Chlamydomonas reinhardtii*. Eukaryotic Cell, 2009. **8**.
396. Valenzuela, J., et al., *Nutrient re-supplementation arrests bio-oil accumulation in Phaeodactylum tricornutum*. Appl Microbiol Biotechnol, 2013. **97**.
397. Breuer, G., et al., *The impact of nitrogen starvation on the dynamics of triacylglycerol accumulation in nine microalgae strains*. Bioresour Technol, 2012. **124**.
398. Wu, Y.H., Y. Yu, and H.Y. Hu, *Potential biomass yield per phosphorus and lipid accumulation property of seven microalgal species*. Bioresour Technol, 2013. **130**: p. 599-602.
399. Li, Y., et al., *Adhesion behavior of marine benthic diatom Nitzschia closterium MMDL533 on cationically modified phosphorylcholine copolymer films*. Asia,Â€Pacific Journal of Chemical Engineering, 2013.
400. Ren, H.Y., et al., *A new lipid-rich microalgae Scenedesmus strain R-16 isolated using Nile Red staining: effects of carbon and nitrogen sources and initial pH on the biomass and lipid production*. Biotechnol Biofuels, 2013. **6**.
401. Kothari, R., et al., *Production of biodiesel from microalgae Chlamydomonas polypyrenoideum grown on dairy industry wastewater*. Bioresour Technol, 2013. **144**.
402. Eustance, E., et al., *Growth, nitrogen utilization and biodiesel potential for two chlorophytes grown on ammonium, nitrate or urea*. Journal of Applied Phycology, 2013: p. 1-15.
403. Bertozzini, E., et al., *Neutral lipid content and biomass production in Skeletonema marinoi (Bacillariophyceae) culture in response to nitrate limitation*. Appl Biochem Biotechnol, 2013. **170**(7): p. 1624-36.
404. Yang, Y., et al., *At high temperature lipid production in Ettlia oleoabundans occurs before nitrate depletion*. Appl Microbiol Biotechnol, 2013. **97**.
405. Feng, P., et al., *Lipid accumulation and growth characteristics of Chlorella zofingiensis under different nitrate and phosphate concentrations*. J Biosci Bioeng, 2012. **114**.
406. Liang, K., et al., *Effect of phosphorus on lipid accumulation in freshwater microalga Chlorella sp.* J Appl Phycol, 2012. **25**.
407. Burrows, E.H., et al., *Dynamics of Lipid Biosynthesis and Redistribution in the Marine Diatom Phaeodactylum tricornutum Under Nitrate Deprivation*. BioEnergy Research, 2012. **5**(4): p. 876-885.
408. Redfield, A., *The biological control of chemical factors in the environment*. Am Sci, 1958. **46**.

409. Palmqvist, K., S. Sjöberg, and G. Samuelsson, *Induction of inorganic carbon accumulation in the unicellular green algae Scenedesmus obliquus and Chlamydomonas reinhardtii*. Plant Physiol, 1988. **87**.
410. Raven, J.A., *Inorganic carbon acquisition by eukaryotic algae: four current questions*. Photosynth Res, 2010. **106**(1-2): p. 123-34.
411. Spalding, M.H., *Microalgal carbon-dioxide-concentrating mechanisms: Chlamydomonas inorganic carbon transporters*. J Exp Bot, 2008. **59**.
412. Sharma, K.K., H. Schuhmann, and P.M. Schenk, *High lipid induction in microalgae for biodiesel production*. Energies, 2012. **5**.
413. Roessler, P.G., *Effects of silicon deficiency in lipid composition and metabolism in the diatom Cyclotella cryptica*. J Phycol, 1988. **24**.
414. Kroger, N. and N. Poulsen, *Diatoms-from cell wall biogenesis to nanotechnology*. Annu Rev Genet, 2008. **42**: p. 83-107.
415. Raven, J.A., *The transport and function of silicon in plants*. Biol Rev, 1983. **58**.
416. Burrows, E.H., et al., *Dynamics of lipid biosynthesis and redistribution in the marine diatom Phaeodactylum tricornutum under nitrate deprivation*. BioEnergy Res, 2012. **5**.
417. Smith, S.R., R.M. Abbriano, and M. Hildebrand, *Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways*. Algal Res, 2012. **1**.
418. Lombardi, A. and P. Wangersky, *Influence of phosphorus and silicon on lipid class production by the marine diatom Chaetoceros gracilis grown in turbidostat cage cultures*. Marine ecology progress series. Oldendorf, 1991. **77**(1): p. 39-47.
419. McGinnis, K., T. Dempster, and M. Sommerfeld, *Characterization of the growth and lipid content of the diatom Chaetoceros muelleri*. Journal of Applied Phycology, 1997. **9**(1): p. 19-24.
420. Obata, T., A.R. Fernie, and A. Nunes-Nesi, *The central carbon and energy metabolism of marine diatoms*. Metabolites, 2013. **3**(2): p. 325-46.
421. Yu, E., et al., *Triacylglycerol accumulation and profiling in the model diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum (Bacillariophyceae) during starvation*. Journal of Applied Phycology, 2009. **21**(6): p. 669-681.
422. Taguchi, S., J.A. Hirata, and E.A. Laws, *Silicate deficiency and lipid synthesis of marine diatoms*. J Phycol, 1987. **23**.
423. Darley, W.M. and B.E. Volcani, *Role of silicon in diatom metabolism: A silicon requirement for deoxyribonucleic acid synthesis in the diatom Cyclindrotheca fusiformis Reimann and Lewin*. Experimental Cell Research, 1969. **58**(2-3): p. 334-342.
424. Buesseler, K.O., et al., *The effects of iron fertilization on carbon sequestration in the southern ocean*. Science, 2004. **304**.
425. Allen, A.E., et al., *Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation*. Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10438-43.
426. Morel, F.M.M., J.G. Rueter, and N.M. Price, *Iron nutrition of phytoplankton and its possible importance in the ecology of ocean regions with high nutrient and low biomass*. Oceanography, 1991. **4**.
427. Johnson, M.B. and Z. Wen, *Development of an attached microalgal growth system for biofuel production*. Appl Microbiol Biotechnol, 2010. **85**.

428. Christenson, L.B. and R.C. Sims, *Rotating algal biofilm reactor and spool harvester for wastewater treatment with biofuels by-products*. Biotechnol Bioeng, 2012. **109**.
429. Schnurr, P.J., G.S. Espie, and D.G. Allen, *Algae biofilm growth and the potential to stimulate lipid accumulation through nutrient starvation*. Bioresour Technol, 2013. **136**: p. 337-44.
430. Ozkan, A., et al., *Reduction of water and energy requirement of algae cultivation using an algae biofilm photobioreactor*. Bioresour Technol, 2012. **114**.
431. Patil, J.S. and A.C. Anil, *Biofilm diatom community structure: influence of temporal and substratum variability*. Biofouling, 2005. **21**.
432. Irving, T.E. and D.G. Allen, *Species and material considerations in the formation and development of microalgal biofilms*. Appl Microbiol Biotechnol, 2011. **92**.
433. Avendaño-Herrera, R.E. and C.E. Riquelme, *Production of a diatom-bacteria biofilm in a photobioreactor for aquaculture applications*. Aquac Eng, 2007. **36**.
434. Tilman, D., *Tests of resource competition theory using 4 species of Lake Michigan algae*. Ecology, 1981. **62**.
435. Downing, A.L. and M.A. Leibold, *Ecosystem consequences of species richness and composition in pond food webs*. Nature, 2002. **416**.
436. Striebel, M., S. Behl, and H. Stibor, *The coupling of biodiversity and productivity in phytoplankton communities: consequences for biomass stoichiometry*. Ecology, 2009. **90**.
437. Interlandi, S.J. and S.S. Kilham, *Limiting resources and the regulation of diversity in phytoplankton communities*. Ecology, 2001. **82**.
438. Stockenreiter, M., et al., *The effect of species diversity on lipid production by microalgal communities*. J Appl Phycol, 2012. **24**.
439. Schindler, D.W., et al., *The cultural eutrophication of Lac la Biche, Alberta, Canada: a paleoecological study*. Can J Fish Aquat Sci, 2008. **65**.
440. Schindler, D.W., *Evolution of phosphorus limitation in lakes*. Science, 1977. **195**.
441. Smith, V.H. and S.J. Bennett, *Nitrogen:phosphorus supply ratios and phytoplankton community structure in lakes*. Arch Hydrobiol, 1999. **146**.
442. Stockenreiter, M., et al., *Functional group richness: implications of biodiversity for light use and lipid yield in microalgae*. J Phycol, 2013. **49**.
443. Tilman, D., *Resource competition between planktonic algae—experimental and theoretical approach*. Ecology, 1977. **58**.
444. Murray, A.G., *Phytoplankton exudation—exploitation of the microbial loop as a defense against algal viruses*. J Plankton Res, 1995. **17**.
445. Rhodes, C.J. and A.P. Martin, *The influence of viral infection on a plankton ecosystem undergoing nutrient enrichment*. J Theo Biol, 2010. **265**.
446. Smith, S., *Organic contaminants in sewage sludge (biosolids) and their significance for agricultural recycling*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2009. **367**(1904): p. 4005-4041.
447. Singh, N.K. and D.W. Dhar, *Microalgae as second generation biofuel*. Agron Sustain Dev, 2011. **31**.
448. Naumann, T., et al., *Growing microalgae as aquaculture feeds on twin-layers: a novel solid-state photobioreactor*. J Appl Phycol, 2013. **25**.
449. Huesemann, M.H., et al., *A screening model to predict microalgae biomass growth in photobioreactors and raceway ponds*. Biotechnol Bioeng, 2013. **110**.

450. Sander, K. and G.S. Murthy, *Life cycle analysis of algae biodiesel*. Int J Life Cycle Assess, 2010. **15**.
451. Clarens, A.F., et al., *Environmental life cycle comparison of algae to other bioenergy feedstocks*. Environ Sci Technol, 2011. **44**.
452. Zaimes, G.G. and V. Khanna, *Environmental sustainability of emerging algal biofuels: a comparative life cycle evaluation of algal biodiesel and renewable diesel*. Environ Prog Sustain Energy, 2013. **32**.
453. Klöpffer, W., *Life cycle assessment*. Environ Sci Pollution Res, 1997. **4**.
454. Brennan, L. and P. Owende, *Biofuels from microalgae—a review of technologies for production, processing, and extractions of biofuels and co-products*. Renew Sust Energy Rev, 2010. **14**.
455. Lardon, L., et al., *Life-cycle assessment of biodiesel production from microalgae*. Environ Sci Technol, 2009. **43**.
456. Kirrolia, A., N.R. Bishnoi, and R. Singh, *Microalgae as a boon for sustainable energy production and its future research and development aspects*. Renew Sustain Energy Rev, 2013. **20**.
457. Davis, R., A. Aden, and P.T. Pienkos, *Techno-economic analysis of autotrophic microalgae for fuel production*. Appl Energy, 2011. **88**.
458. Quinn, J.C., et al., *Microalgae to biofuels lifecycle assessment—multiple pathway evaluation*. Bioenergy Res, 2013. **6**.
459. Chowdhury, R., S. Viamajala, and R. Gerlach, *Reduction of environmental and energy footprint of microalgal biodiesel production through material and energy integration*. Bioresour Technol, 2012. **108**: p. 102-11.
460. Torres, C.M., et al., *Microalgae-based biodiesel: a multicriteria analysis of the production process using realistic scenarios*. Bioresour Technol, 2013. **147**.
461. Ríos, S.D., et al., *Microalgae-based biodiesel: economic analysis of downstream process realistic scenarios*. Bioresour Technol, 2013. **136**.
462. Rawat, I., et al., *Biodiesel from microalgae: a critical evaluation from laboratory to large-scale production*. Appl Energy, 2013. **103**.
463. Nagarajan, S., et al., *An updated comprehensive techno-economic analysis of algae biodiesel*. Bioresour Technol, 2013. **145**.
464. Liu, X., et al., *Pilot-scale data provide enhanced estimates of the life cycle energy and emissions profile of algae biofuels produced via hydrothermal liquefaction*. Bioresour Technol, 2013. **148**.
465. Frank, E.D., et al., *Life cycle comparison of hydrothermal liquefaction and lipid extraction pathways to renewable diesel from algae*. Mitig Adapt Strateg Glob Chang, 2013. **18**.
466. Elliott, D.C., G.G. Neuenschwander, and T.R. Hart, *Hydroprocessing bio-oil and products separation for coke production*. Acs Sustainable Chemistry & Engineering, 2013. **1**(4): p. 389-392.
467. López Barreiro, D.L., et al., *Hydrothermal liquefaction (HTL) of microalgae for biofuel production: state of the art review and future prospects*. Biomass Bioenergy, 2013. **53**.
468. Collet, P., et al., *Biodiesel from microalgae—Life cycle assessment and recommendations for potential improvements*. Renewable Energy, 2014. **71**: p. 525-533.

469. Christenson, L. and R. Sims, *Production and harvesting of microalgae for wastewater treatment, biofuels, and bioproducts*. Biotechnology advances, 2011. **29**(6): p. 686-702.
470. Ördög, V., et al., *Screening microalgae for some potentially useful agricultural and pharmaceutical secondary metabolites*. Journal of applied phycology, 2004. **16**(4): p. 309-314.
471. Pokoo-Aikins, G., et al., *Design and analysis of biodiesel production from algae grown through carbon sequestration*. Clean Technologies and Environmental Policy, 2010. **12**(3): p. 239-254.
472. Hall-Stoodley, L., J.W. Costerton, and P. Stoodley, *Bacterial biofilms: from the natural environment to infectious diseases*. Nature reviews microbiology, 2004. **2**(2): p. 95-108.
473. Stewart, P.S. and M.J. Franklin, *Physiological heterogeneity in biofilms*. Nat Rev Microbiol, 2008. **6**(3): p. 199-210.
474. Kühl, M., et al., *Microenvironmental control of photosynthesis and photosynthesis-coupled respiration in an epilithic cyanobacterial biofilm*. Journal of Phycology, 1996. **32**(5): p. 799-812.
475. Falkowski, P.G., and Raven J.A., *Aquatic photosynthesis*. 1997, Blackwell Science: Malden, MA.
476. Boelee, N.C., Temmink, Hardy, Janssen, Marcel, Buisman, Cees J. N. & Wiffels, Rene H., *Scenario Analysis of Nutrient Removal from Municipal Wastewater by Microalgal Biofilms*. Water, 2012. **4**(2): p. 460-473.
477. Revsbech, N.P., *An oxygen microsensor with a guard cathode*. Limnology and Oceanography, 1989. **34**(2): p. 474-478.
478. Stewart, P.S., *A review of experimental measurements of effective diffusive permeabilities and effective diffusion coefficients in biofilms*. Biotechnology and Bioengineering, 1998. **59**(3): p. 261-272.
479. Bernstein, H.C., et al., *In situ analysis of oxygen consumption and diffusive transport in high-temperature acidic iron-oxide microbial mats*. Environmental microbiology, 2013. **15**(8): p. 2360-2370.
480. Glud, R.N., N.B. Ramsing, and N.P. Revsbech, *Photosynthesis and photosynthesis-coupled respiration in natural biofilms quantified with oxygen microsensors*. Journal of Phycology, 1992. **28**(1): p. 51-60.
481. Eustance, E., et al., *Growth, nitrogen utilization, and biodiesel potential for two chlorophytes grown on ammonium, nitrate, or urea*. J Appl Phycol, 2013. **25**.
482. Jørgensen, B.B. and D.J. Des Marais, *The diffusive boundary layer of sediments: oxygen microgradients over a microbial mat*. Limnology and Oceanography, 1990. **35**(6): p. 1343-1355.
483. Kliphuis, A.M., et al., *Effect of O₂: CO₂ ratio on the primary metabolism of Chlamydomonas reinhardtii*. Biotechnology and bioengineering, 2011. **108**(10): p. 2390-2402.
484. Jensen, J. and N.P. Revsbech, *Photosynthesis and respiration of a diatom biofilm cultured in a new gradient growth chamber*. FEMS Microbiology Ecology, 1989. **5**(1): p. 29-38.
485. Converti, A., et al., *Effect of temperature and nitrogen concentration on the growth and lipid content of Nannochloropsis oculata and Chlorella vulgaris for biodiesel production*.

- Chemical Engineering and Processing: Process Intensification, 2009. **48**(6): p. 1146-1151.
486. Stephenson, A.L., et al., *Influence of nitrogen-limitation regime on the production by *Chlorella vulgaris* of lipids for biodiesel feedstocks*. Biofuels, 2010. **1**(1): p. 47-58.
 487. Bernstein, H.C. and R.P. Carlson, *Microbial Consortia Engineering for Cellular Factories: in vitro to in silico systems*. Comput Struct Biotechnol J, 2012. **3**: p. e201210017.
 488. Ellis, J.T., et al., *Acetone, butanol, and ethanol production from wastewater algae*. Bioresource technology, 2012. **111**: p. 491-495.
 489. Gardner, R., et al., *Medium pH and nitrate concentration effects on accumulation of triacylglycerol in two members of the chlorophyta*. J Appl Phycol, 2011. **23**.
 490. Gardner, R., et al., *Use of sodium bicarbonate to stimulate triacylglycerol accumulation in the chlorophyte *Scenedesmus* sp. and the diatom *Phaeodactylum tricornutum**. J Appl Phycol, 2012. **24**.
 491. Cai, T., S.Y. Park, and Y. Li, *Nutrient recovery from wastewater streams by microalgae: status and prospects*. Renewable and Sustainable Energy Reviews, 2013. **19**: p. 360-369.
 492. Sturm, B.S. and S.L. Lamer, *An energy evaluation of coupling nutrient removal from wastewater with algal biomass production*. Applied Energy, 2011. **88**(10): p. 3499-3506.
 493. Hoffmann, J.P., *Wastewater treatment with suspended and nonsuspended algae*. Journal of phycology, 1998. **34**(5): p. 757-763.
 494. Gross, M., *The mysteries of the diatoms*. Curr Biol, 2012. **22**(15): p. R581-5.
 495. Kesaano, M., et al., *Dissolved inorganic carbon enhanced growth, nutrient uptake, and lipid accumulation in wastewater grown microalgal biofilms*. Bioresour Technol, 2015. **180C**: p. 7-15.
 496. Pizarro, C., et al., *An economic assessment of algal turf scrubber technology for treatment of dairy manure effluent*. Ecological Engineering, 2006. **26**(4): p. 321-327.
 497. Su, C.-H., et al., *Factors affecting lipid accumulation by *Nannochloropsis oculata* in a two-stage cultivation process*. Journal of Applied Phycology, 2011. **23**(5): p. 903-908.
 498. Devi, M.P., G.V. Subhash, and S.V. Mohan, *Heterotrophic cultivation of mixed microalgae for lipid accumulation and wastewater treatment during sequential growth and starvation phases: effect of nutrient supplementation*. Renewable energy, 2012. **43**: p. 276-283.
 499. Rodolfi, L., et al., *Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor*. Biotechnol Bioeng, 2009. **102**(1): p. 100-12.
 500. Peng, X., et al., *Triacylglycerol accumulation of *Phaeodactylum tricornutum* with different supply of inorganic carbon*. Journal of applied phycology, 2014. **26**(1): p. 131-139.
 501. White, D., et al., *The effect of sodium bicarbonate supplementation on growth and biochemical composition of marine microalgae cultures*. Journal of Applied Phycology, 2013. **25**(1): p. 153-165.
 502. Chi, Z., et al., *Bicarbonate-based integrated carbon capture and algae production system with alkalihalophilic cyanobacterium*. Bioresource technology, 2013. **133**: p. 513-521.
 503. Metcalf, E., *Wastewater Engineering: Treatment and Reuse*. fourth ed ed. 2003, New York: McGraw Hill.

- 504. Rhine, E., et al., *Improving the Berthelot reaction for determining ammonium in soil extracts and water*. Soil Science Society of America Journal, 1998. **62**(2): p. 473-480.
- 505. Crofcheck, C.L., et al. *Influence of media composition on the growth rate of Chlorella vulgaris and Scenedesmus acutus utilized for CO2 mitigation*. in *2012 Dallas, Texas, July 29-August 1, 2012*. 2012. American Society of Agricultural and Biological Engineers.
- 506. Glud, R.N., *Oxygen dynamics of marine sediments*. Marine Biology Research, 2008. **4**(4): p. 243-289.
- 507. Wieland, A. and M. Kühl, *Irradiance and temperature regulation of oxygenic photosynthesis and O2 consumption in a hypersaline cyanobacterial mat (Solar Lake, Egypt)*. Marine Biology, 2000. **137**(1): p. 71-85.
- 508. Mus, F., et al., *Physiological and molecular analysis of carbon source supplementation and pH stress-induced lipid accumulation in the marine diatom Phaeodactylum tricornutum*. Appl Microbiol Biotechnol, 2013. **97**.

APPENDICES

APPENDIX A

CHARACTERIZATION OF NINE NOVEL GREEN ALGAE STRAINS FROM
YELLOWSTONE NATIONAL PARK

Supplementary Data

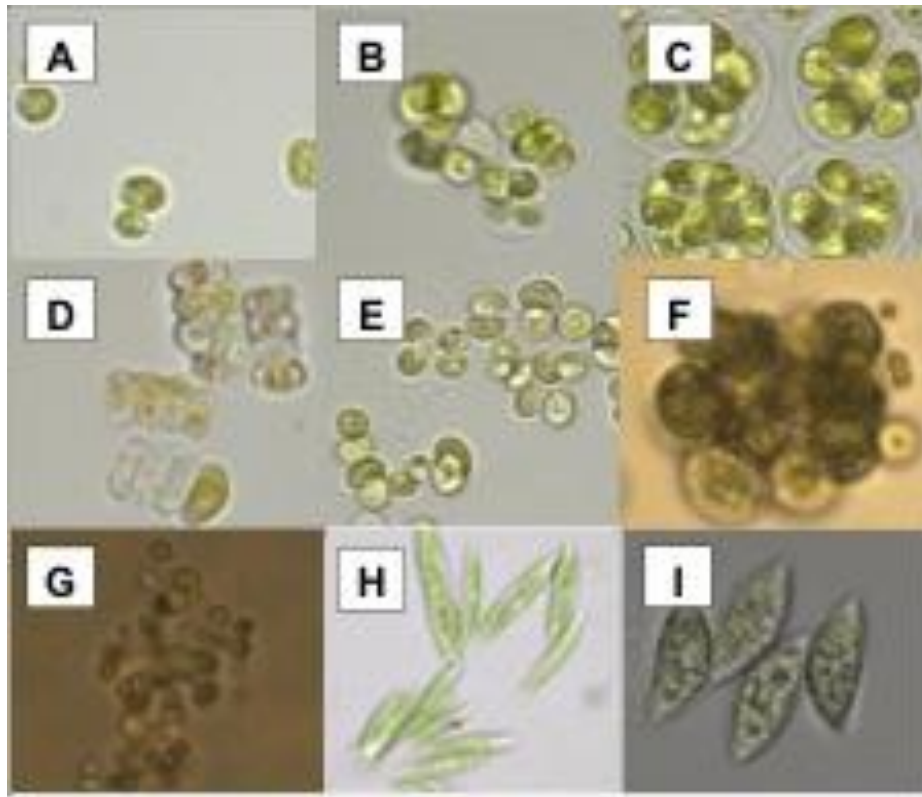


Figure A.1 Light microscopy images of the nine YNP green algae isolates. (A) PGV-6 (B) PGV8-G1 (C) PGV8-G2 (D) PGV10-G1 (E) PGV10-G2 (F) WC2b (G) WC-5A (H) MF1 and (I) WC-1.

Table A.1 The optimal Nile Red exposure stain times and stain methods for each of the 11 green algae strains. Each strain was exposed to the lipophilic stain, Nile Red, in 20% DMSO and acetone until an optimal stain time was indicated. The stain method that resulted in higher fluorescence was selected as the proper stain method for each strain because that carrier (DMSO or acetone) was able to cross the cell membrane more effectively.¹

Strain	Stain Time	Stain Method
MF1	60 min	20% DMSO
PGV-6	60 min	20% DMSO
PGV8-G1	4 min	20% DMSO
PGV8-G2	6 min	20% DMSO
PGV10-G1	60 min	20% DMSO
PGV10-G2	40 min	20% DMSO
WC-1	60 min	Acetone
WC-2B	10 min	Acetone
WC-5A	4 min	20% DMSO
PC-3A	30 min	Acetone
UTEX-395	10 min	20% DMSO

Table A.2 The endpoint DCW and doubling times in the air-only and sodium bicarbonate added conditions for each of the 11 strains. Each condition was grown in triplicate.

	Dry Cell Weight (g/L)				Doubling time (h)			
	air only		air + HCO ₃ ⁻		air only		air + HCO ₃ ⁻	
	Ave	Std dev	Ave	Std dev	Ave	Std dev	Ave	Std dev
MF1	1.00	0.06	1.22	0.11	24.79	0.91	25.79	3.15
PC3	1.08	0.22	1.50	0.12	27.45	4.20	26.44	1.56
PGV10-G1	1.02	0.08	0.69	0.09	24.33	0.32	26.23	1.36
PGV10-G2	0.95	0.13	0.95	0.19	21.37	1.44	24.15	1.92
PGV6	1.06	0.26	1.08	0.06	30.40	2.51	29.76	2.44
PGV8-G1	0.56	0.10	0.65	0.06	24.63	0.18	20.39	0.19
PGV8-G2	0.73	0.12	0.54	0.01	17.36	5.64	22.62	7.77
UTEX 395	0.83	0.02	1.03	0.12	25.08	2.14	27.44	1.78
WC-1	1.02	0.06	1.18	0.14	19.80	6.35	16.07	1.28
WC-2b	1.02	0.14	1.44	0.08	18.13	0.63	15.00	0.43
WC-5	0.86	0.05	0.91	0.17	15.98	1.31	20.97	1.41

Table A.3 Final DCWs and doubling times for each green algae strain for the control and sodium bicarbonate addition conditions. The DCWs were the average and 95% confidence interval of each triplicate at the time of harvest for each experiment.

Strain	50mM NaHCO ₃	Cell Dry Weight [g·L ⁻¹]
WC-1	Control	1.02 ± 0.06
	Bicarbonate	1.18 ± 0.14
WC-2b	Control	1.02 ± 0.14
	Bicarbonate	1.44 ± 0.08
WC-5	Control	0.86 ± 0.05
	Bicarbonate	0.91 ± 0.17
PGV-6	Control	1.06 ± 0.26
	Bicarbonate	1.08 ± 0.06
PGV-8 G1	Control	0.56 ± 0.10
	Bicarbonate	0.65 ± 0.06
PGV-8 G2	Control	0.73 ± 0.12
	Bicarbonate	0.54 ± 0.01
PGV-10 G1	Control	1.02 ± 0.08
	Bicarbonate	0.69 ± 0.09
PGV-10 G2	Control	0.95 ± 0.13
	Bicarbonate	0.95 ± 0.18
MF1	Control	1.00 ± 0.06
	Bicarbonate	1.22 ± 0.11
PC-3a	Control	1.08 ± 0.22
	Bicarbonate	1.50 ± 0.12
UTEX 395	Control	0.83 ± 0.02
	Bicarbonate	1.02 ± 0.12

APPENDIX B

RGD-1 GENOME SUPPLEMENTARY DATA

Hight Molecular Weight DNA Extraction

Table 4 DNA Extraction (JGI Method).² The 1.5, 30 and 60 mL headers refer to the container volumes recommended for the DNA extraction volumes. Fifty milliliters were centrifuged and adjusted to ~OD₆₀₀ 1.0 as indicated in step 6. To improve cell wall breakage, mechanical stress was applied with sterile sand, mortar and pestle and, liquid nitrogen. Rather than using isopropanol, the DNA was suspended in molecular grade ethanol in the -20°C freezer overnight to improve DNA precipitation.

	<u>1.5ml</u>	<u>30ml</u>	<u>60ml</u>
1. Grow cells (see above) in broth and pellet at 10,000 rpm for 5 min or scrape from plate.			
2. Transfer bacterial suspension to the appropriate centrifuge tube.			
3. Spin down cells in microfuge or centrifuge at 10,000 rpm for 5 minute.			
4. Discard the supernatant.			
5. Resuspend cells in TE.			
6. Adjust to OD ⁶⁰⁰ $\cong$ 1.0 with TE buffer (10mM tris; 1 mM EDTA, pH 8.0)			
7. Transfer given amount of cell suspension to a clean centrifuge tube. -----	740 μ l	14.8ml	29.6ml
8. Add lysozyme (conc. 100mg/ml). Mix well. ----- This step is necessary for hard to lyse gram (+) and some gram (-) bacteria.	20 μ l	400 μ l	800 μ l
9. Incubate for 5 min. at room temperature.			
10. Add 10% SDS. Mix well. -----	40 μ l	800 μ l	1.6ml
11. Add Proteinase K (10mg/ml). Mix well. -----	8 μ l	160 μ l	320 μ l
12. Incubate for 1 hr at 37°C.			
13. Add 5 M NaCl. Mix well. -----	100 μ l	2ml	4ml
14. Add CTAB/NaCl (heated to 65°C). Mix well. -----	100 μ l	2ml	4ml
15. Incubate at 65°C for 10 min.			
16. Add chloroform:isoamyl alcohol (24:1). Mix well. -----	0.5ml	10ml	20ml
17. Spin at max speed for 10 min at room temperature.			
18. Transfer aqueous phase to clean eppendorf (should not be viscous).			
19. Add phenol:chloroform:isoamyl alcohol (25:24:1). Mix well. -----	0.5ml	10ml	20ml
20. Spin at max speed for 10 min at room temperature.			
21. Transfer aqueous phase and add 0.6 vol isopropanol (-20°C). (e.g. if 400 μ l of aqueous phase is transferred, add 240 μ l of isopropanol.			
	----	Add 0.6 of vol.	----
22. Incubate at room temp for 30 min.			
23. Spin at max speed for 15 min.			
24. Wash pellet with 70% ethanol, spin at max speed for 5 min.			
25. Discard the supernatant and and let pellet dry for 5 – 10 min at room temp.			
26. Resuspend in TE plus RNase (99 μ l TE + 1 μ l RNase (10 mg/ml)). -----	20 μ l	400 μ l	800 μ l
27. Transfer to sterile microcentrifuge tubes.			
28. Incubate at 37°C for 20 min.			
29. Run 1 μ l in a 1% agarose gel with concentration standards.			

BioNano and Assembly Data

Table B.1 Assembly statistics for BioNano data. RGD-1 biomass was submitted to the Bioinformatics Center at Kansas State for high molecular weight (HMW) DNA extraction and whole genome map assembly. The HMW DNA was digested using the endonuclease, Dnase1 to introduce nicks and create 3' hydroxyl group. DNA polymerase 1 catalyzed the addition of fluorescently labeled Alexa 546 dUTP fluorescent dyes that attached to the nucleotides at the 3' hydroxyl group. 5' to 3' exonuclease activity removed the nucleotides from the 5' phosphoryl terminus of the nick. The labeled and unlabeled nucleotides displaced the excised nucleotides in the original DNA strand. The fluorescently-labeled DNA were visualized using the intercalating dye, YOYO-1. The labeled DNA was added to an IrysChip flow cell, linearized with an electrophoretic current and imaged.

Assembly Name	N conti gs	Total Conti g Len(M b)	Avg. Conti g Len(Mb)	Conti g N50(Mb)	Total Ref Len(Mb)	Total Conti g Len / Ref Len	N conti gs total align I	Total Align ed Len(Mb) I	Total Align ed Len / Ref Len I	Total Uniq ue Align ed Len(Mb) I	Total Uniq ue Len / Ref Len I	N conti gs total align II	Total Align ed Len(Mb) II	Total Align ed Len / Ref Len II	Total Uniq ue Align ed Len(Mb) Final	Total Uniq ue Len / Ref Len II
Sella_pupl_0031: strict_t	57	20.144	0.353	0.403	7.396	2.724	0 (0.00)	0	0	0	0	2 (0.04)	0.096	0.013	0.096	0.013
Sella_pupl_0031: default_t	52	21.844	0.42	0.579	7.396	2.953	0 (0.00)	0	0	0	0	3 (0.06)	0.124	0.017	0.124	0.017
Sella_pupl_0031: relaxed_t	53	22.376	0.422	0.602	7.396	3.025	0 (0.00)	0	0	0	0	2 (0.04)	0.096	0.013	0.096	0.013

Table B.2 Pfam proteins. Seven Pfam proteins were found in common among the 18 algal genomes that were used for the concatenated protein tree using the ezTree, pipeline.

Pfam ID		
PF08149.10	BING4CT	BING4CT
PF12295.7	Symplekin_C	Symplekin
PF03332.12	PMM	Eukaryotic
PF04034.12	Ribo_biogen_C	Ribosome
PF00692.18	dUTPase	dUTPase
PF06862.11	UTP25	Utp25
PF03568.16	Peptidase_C50	Peptidase

Transcript Raw Read Data

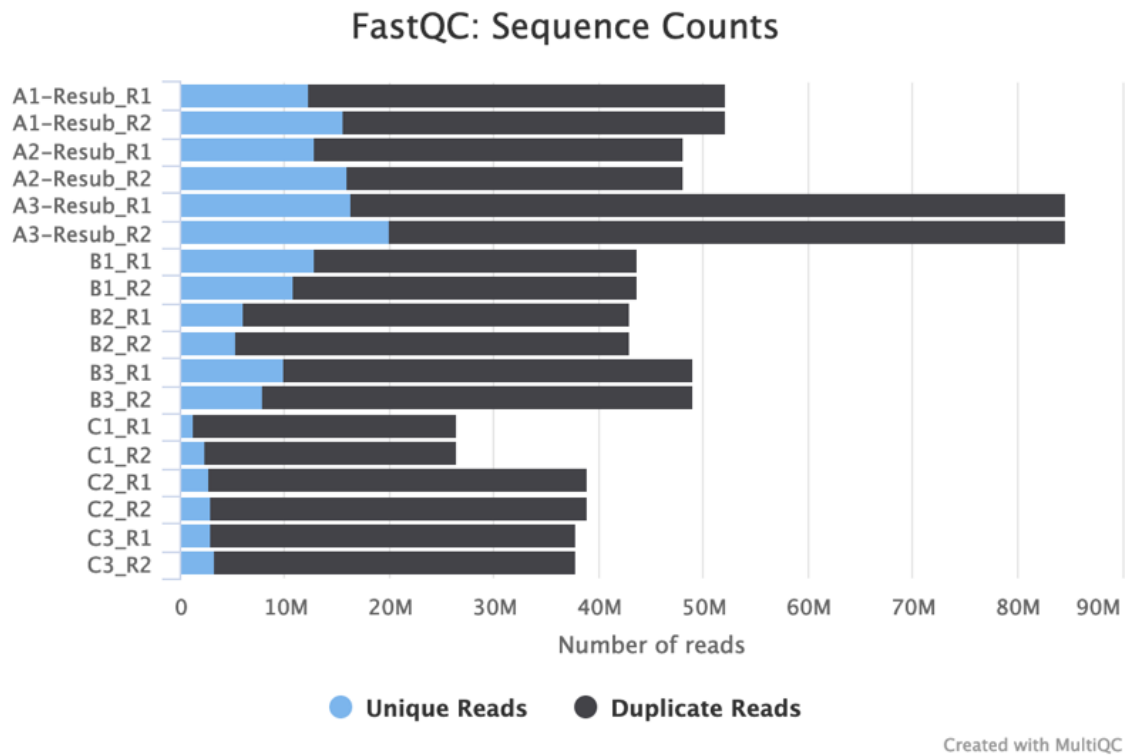


Figure B.1 Each sample was analyzed using FastQC and compiled within MultiQC for their unique and duplicate sequence counts. There were a total of nine samples, three culture conditions and three replicates for each condition. The forward (R1) and reverse (R2) reads were analyzed for each sample. Samples A1-A3 had the largest number of unique reads among the sequenced samples and C1-C3 had the least number of unique reads.

Sample A1-R1

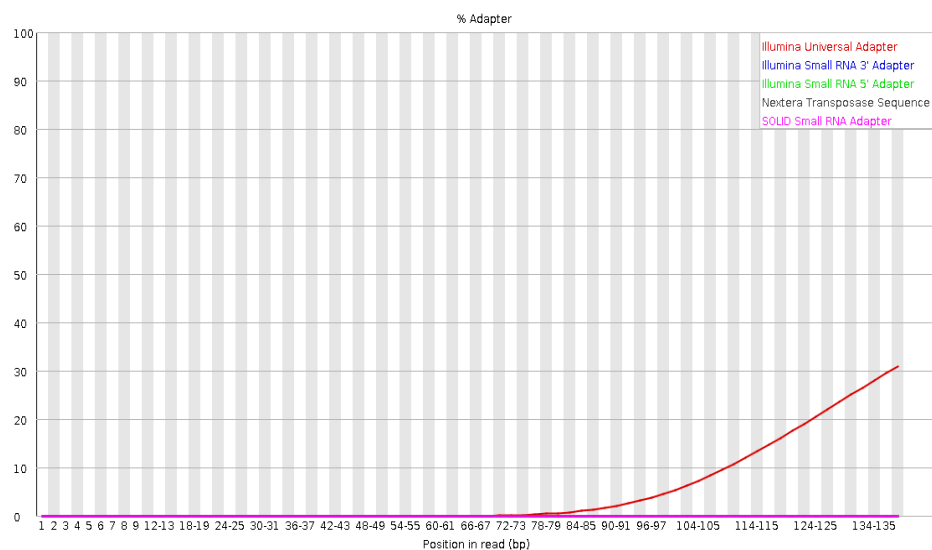


Figure B.2 This figure indicates the presence of adapter sequence contamination in sample A1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

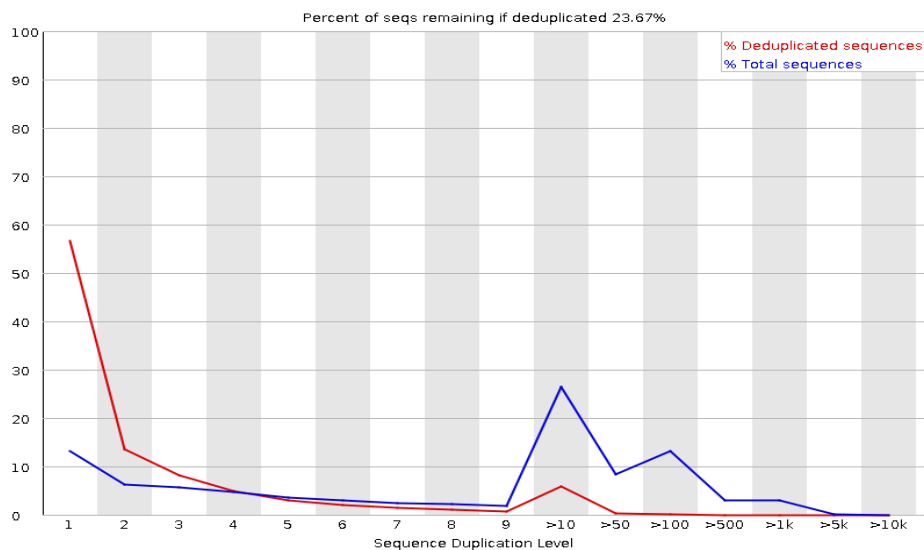


Figure B.2 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A1-R1. Here, 23.67% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

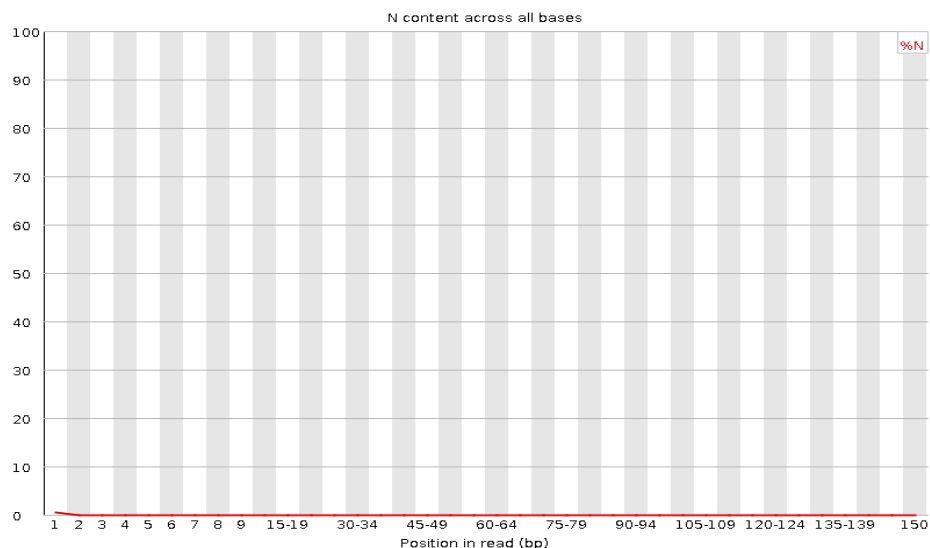


Figure B.4 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

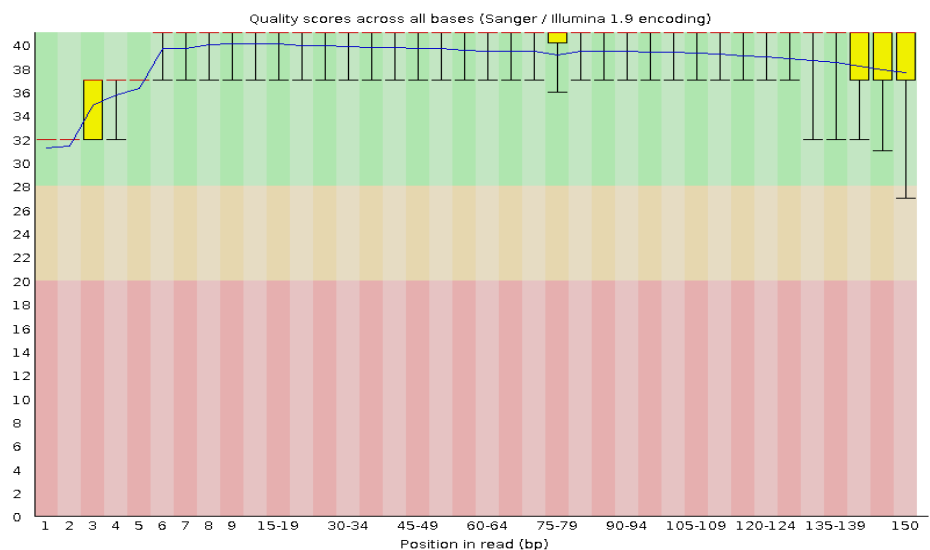


Figure 2 Quality scores across the positions of the 150 bp reads for sample A1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

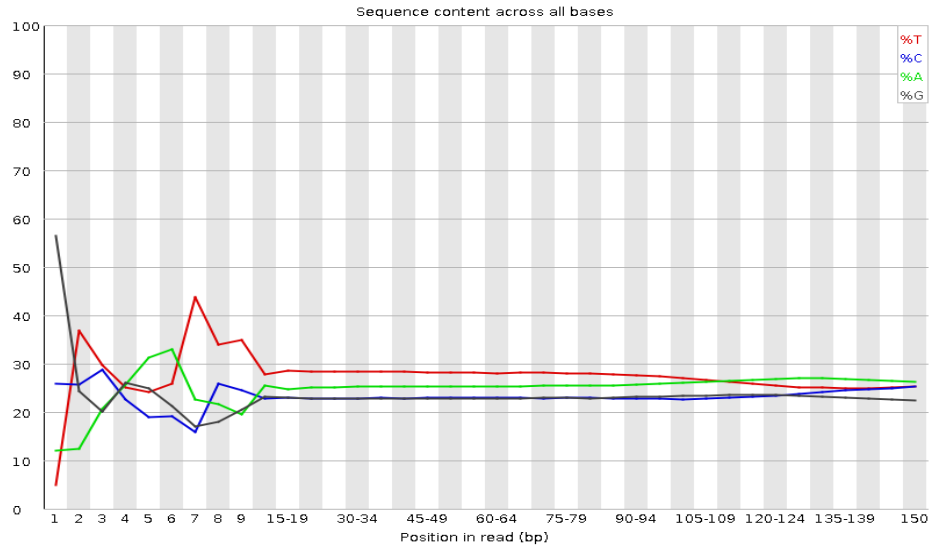


Figure B.6 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

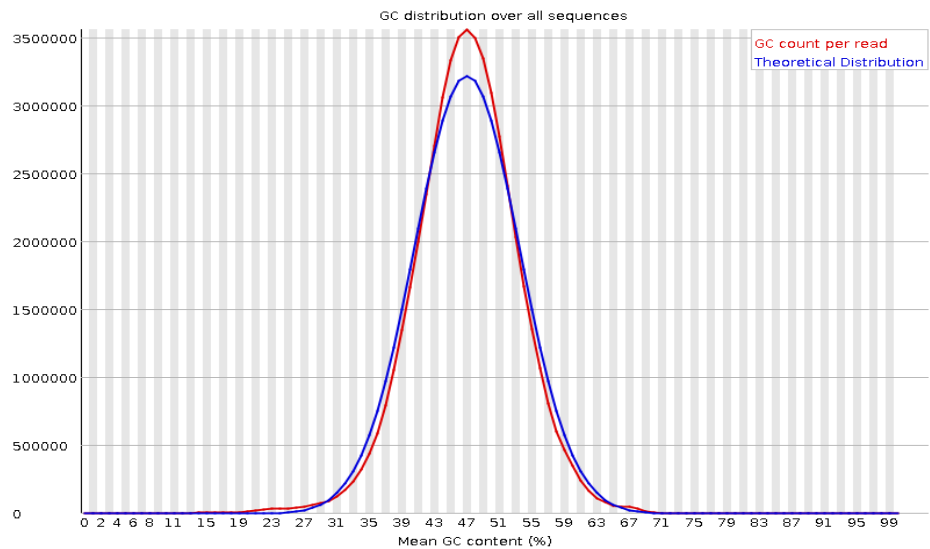


Figure B.7 The per sequence GC content for sample A1-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

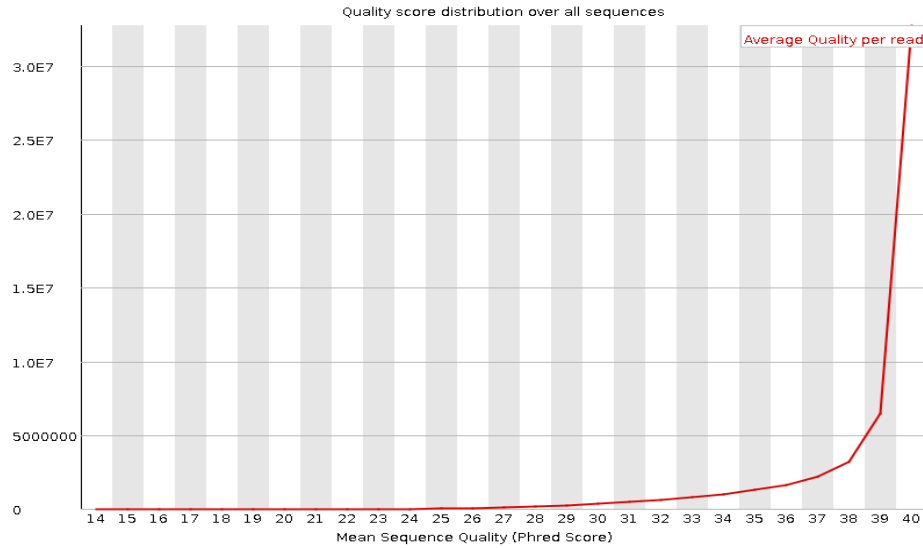


Figure B.8 The per sequence quality score for sample A1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.9 The distribution of the sequence lengths for sample A1-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample A1-R2

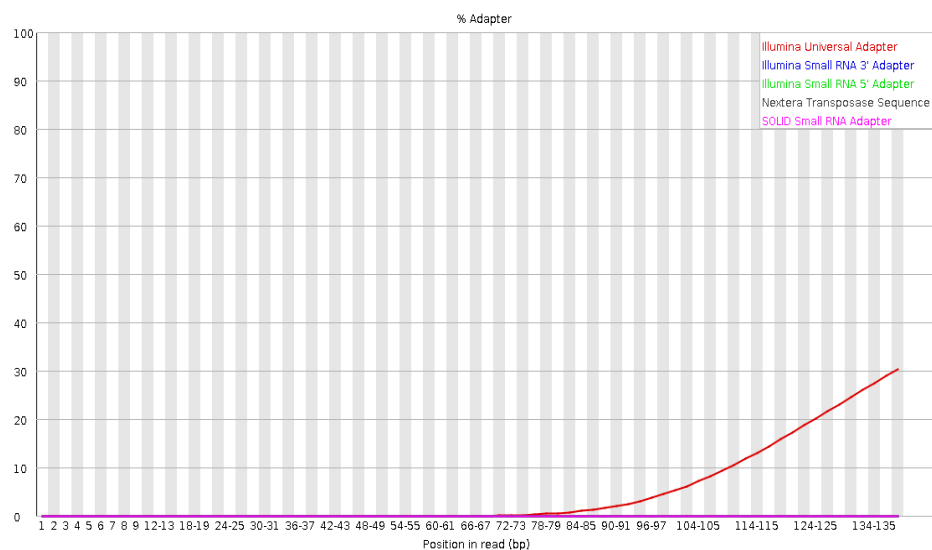


Figure B.10 This figure indicates the presence of adapter sequence contamination in sample A1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

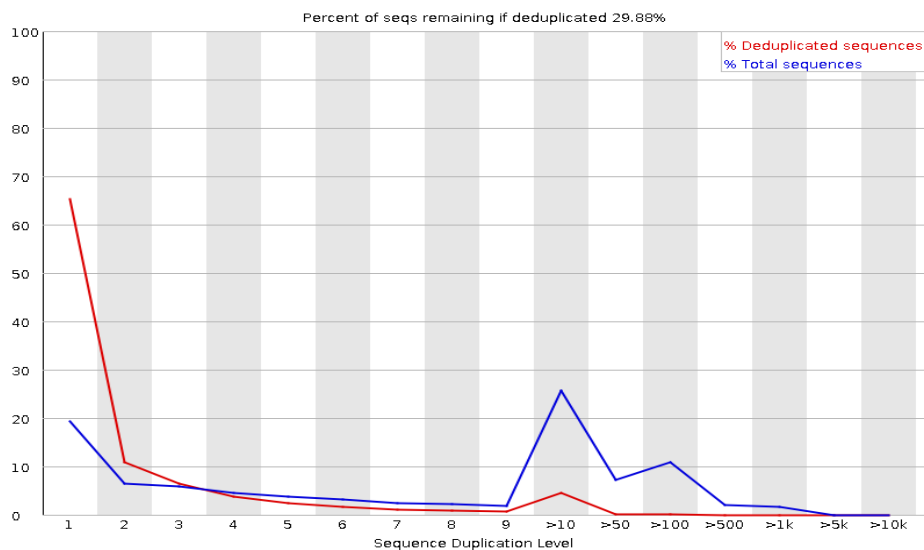


Figure B.11 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A1-R2. Here, 29.88% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

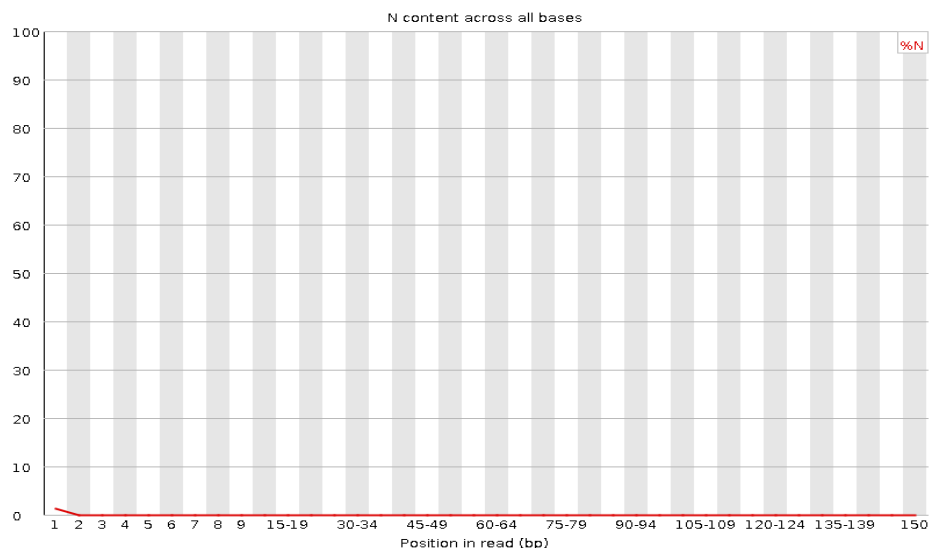


Figure B.12 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

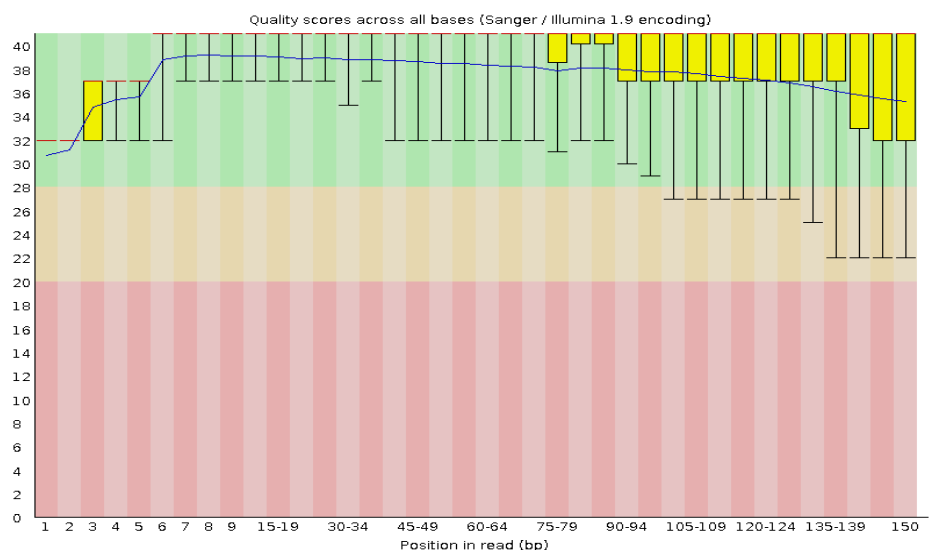


Figure B.13 Quality scores across the positions of the 150 bp reads for sample A1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

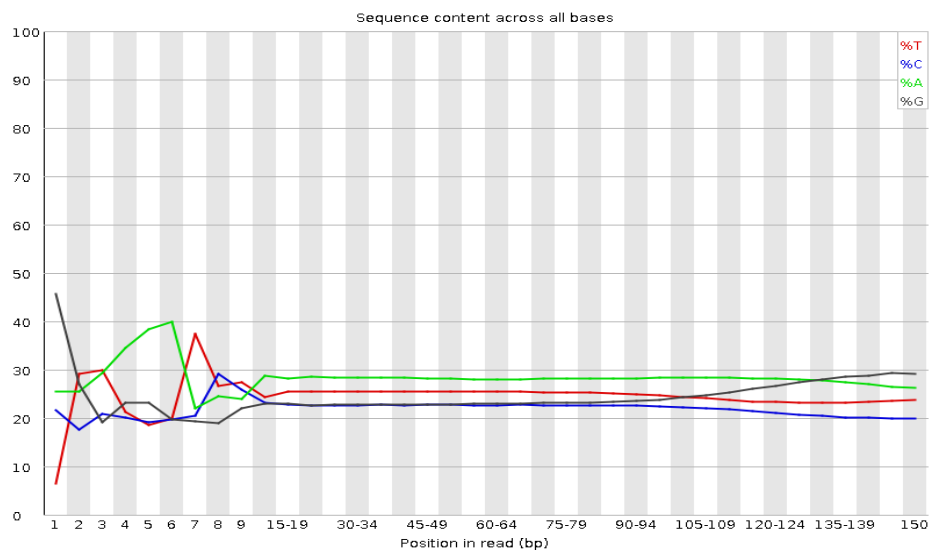


Figure B.14 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

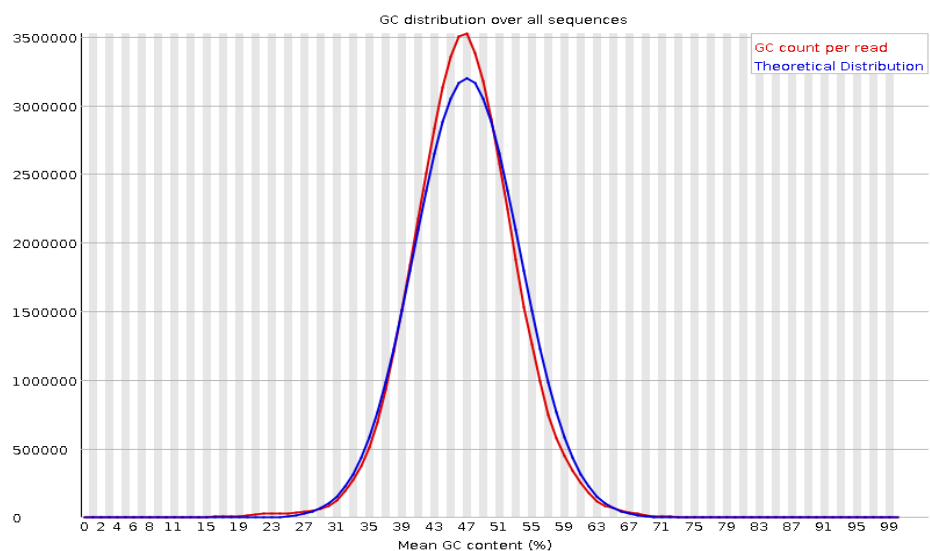


Figure B.15 The per sequence GC content. The x- and y-axes represent the %GC content per read and read counts, respectively for sample A1-R2. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

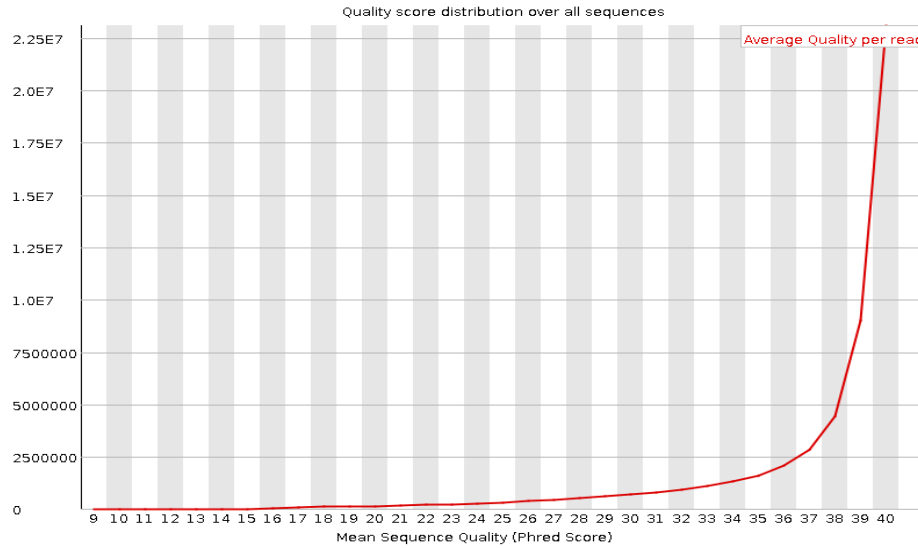


Figure B.16 The per sequence quality score for sample A1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .

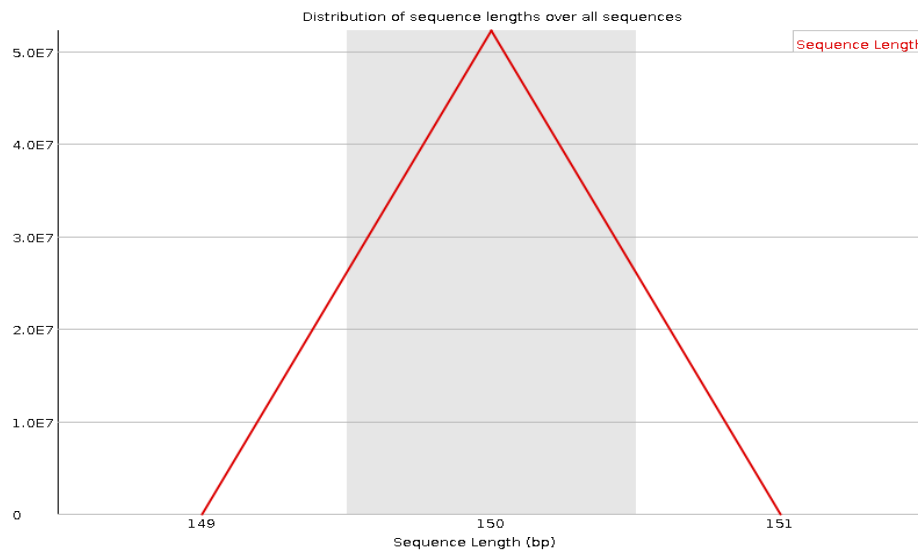


Figure B.17 The distribution of the sequence lengths for sample A1-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample A2-R1

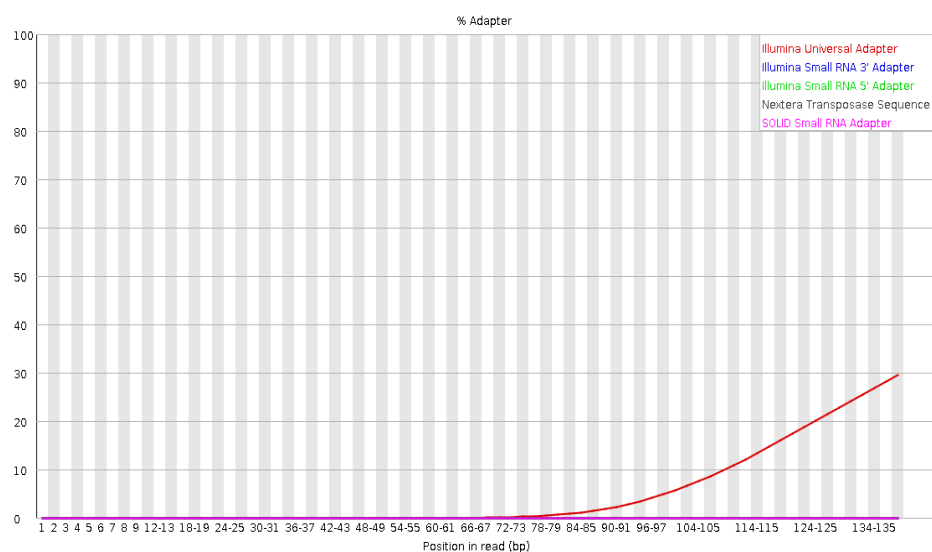


Figure B.18 This figure indicates the presence of adapter sequence contamination in sample A2-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

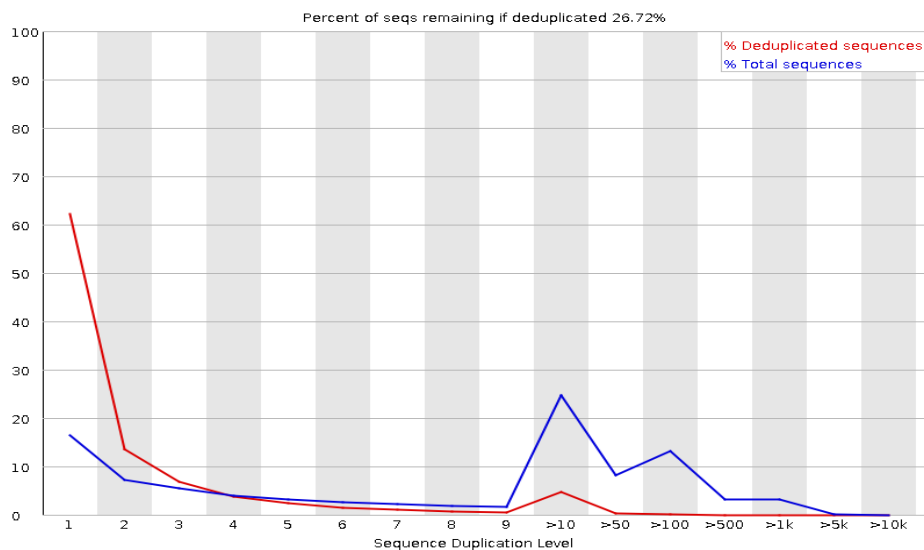


Figure B.19 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A2-R1. Here, 26.72% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

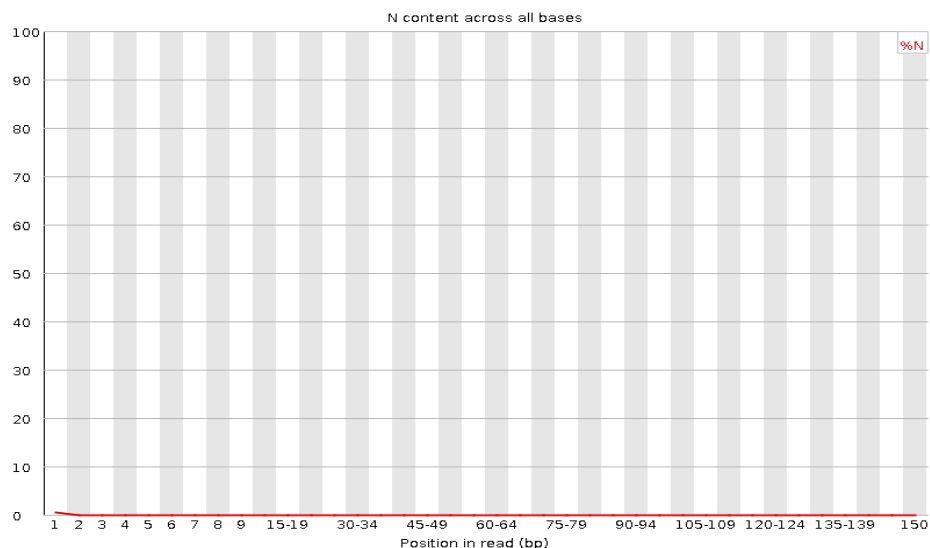


Figure B.20 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A2-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

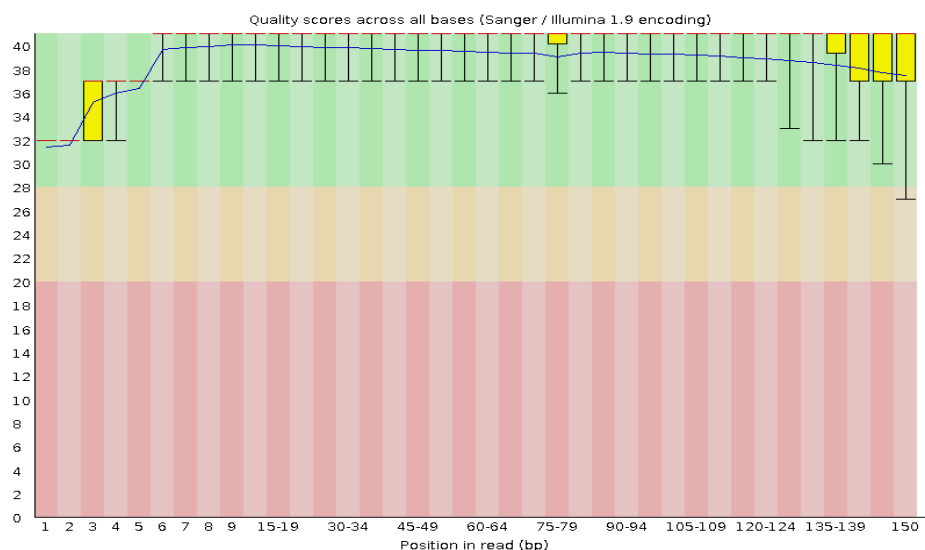


Figure B.21 Quality scores across the positions of the 150 bp reads for sample A2-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

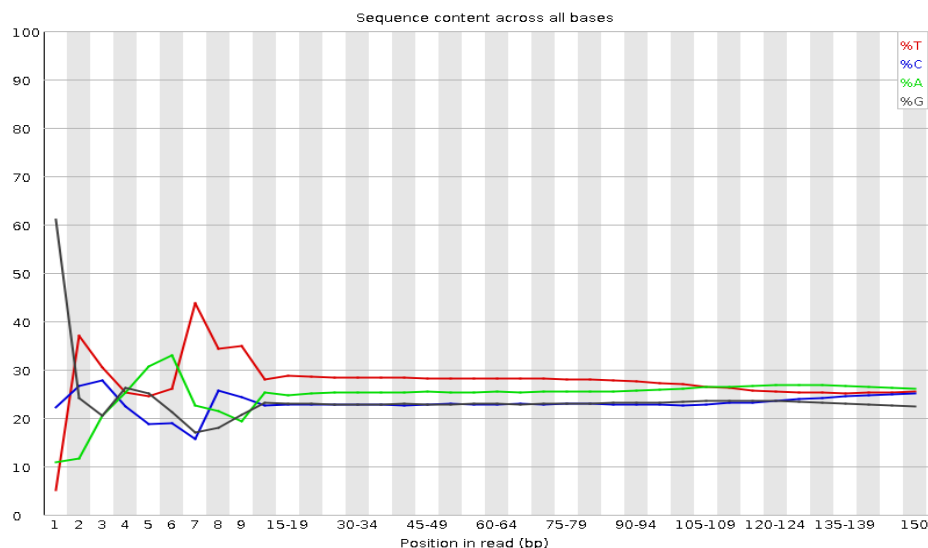


Figure B.22 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A2-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

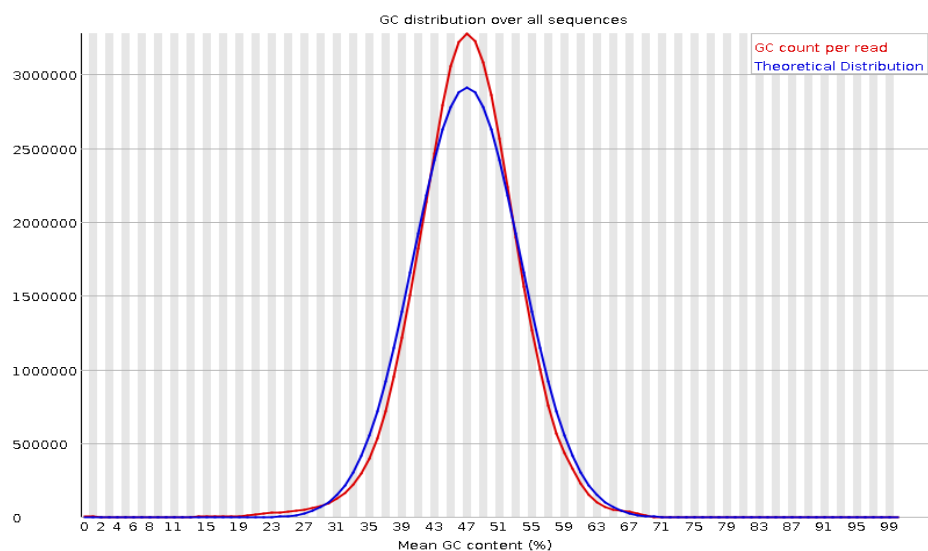


Figure B.23 The per sequence GC content for sample A2-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

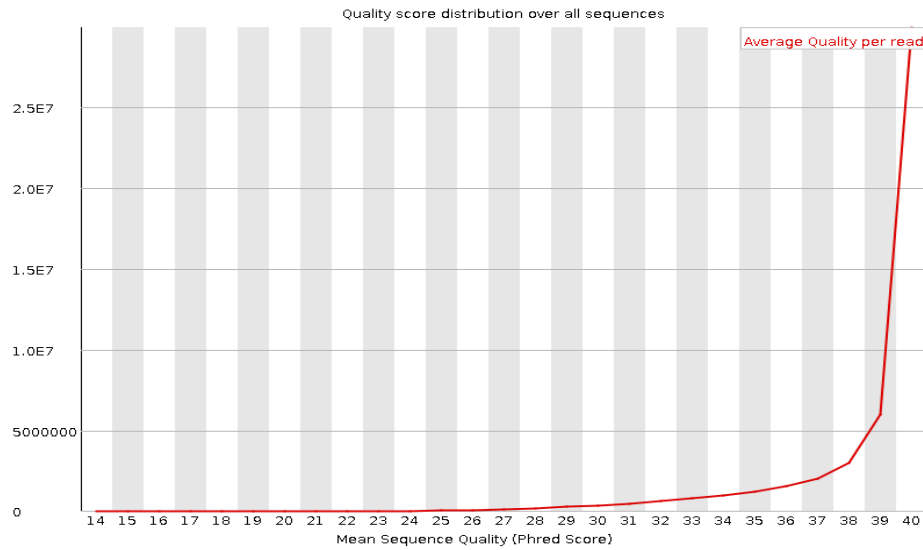


Figure B.24 The per sequence quality score for sample A2-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.

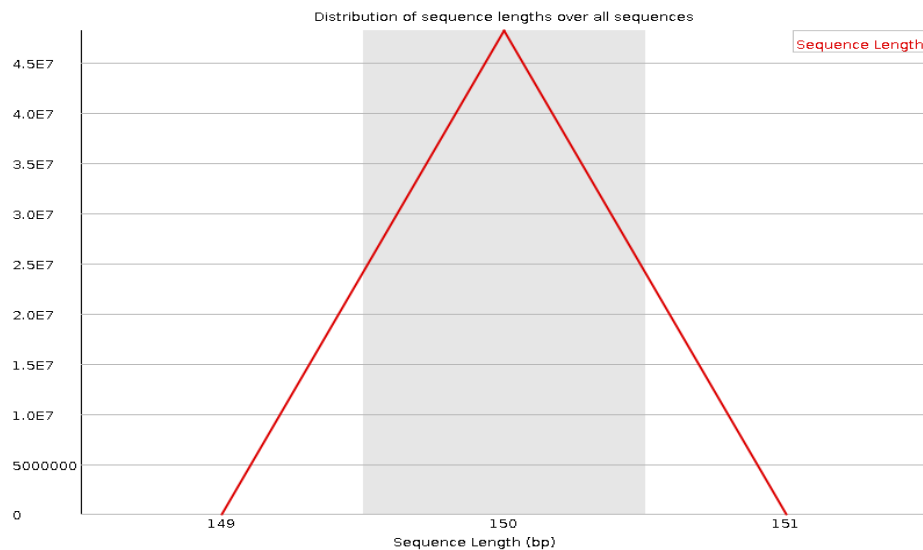


Figure B.25 The distribution of the sequence lengths for sample A2-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample A2-R2

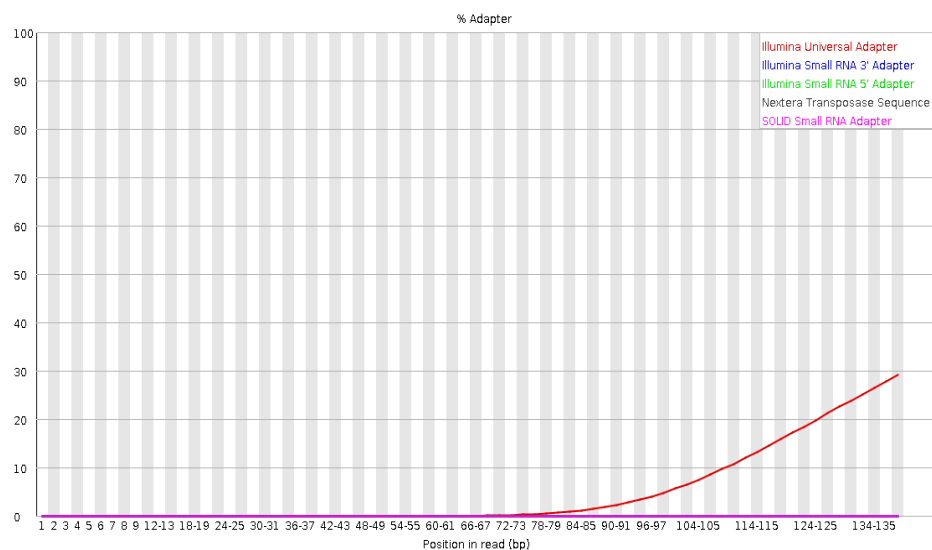


Figure B.26 This figure indicates the presence of adapter sequence contamination in sample A2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

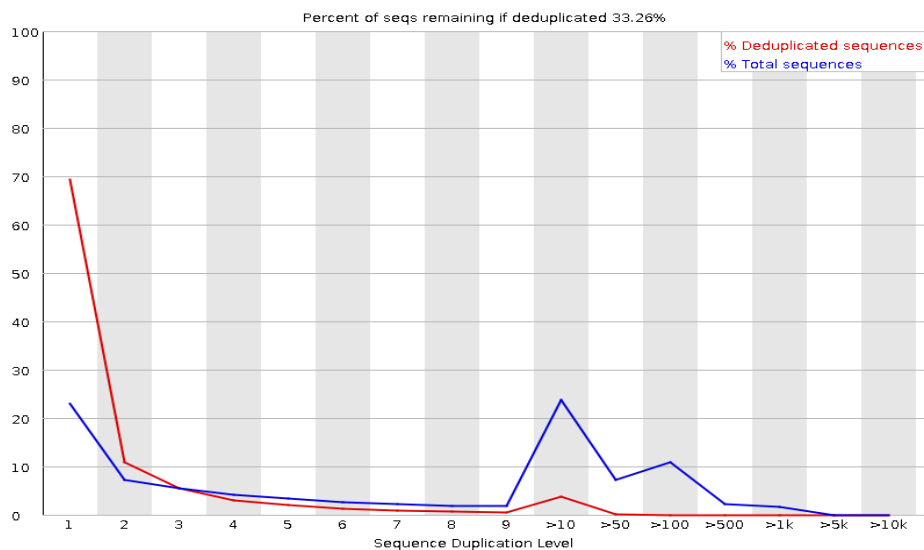


Figure B.27 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A2-R2. Here, 33.26% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

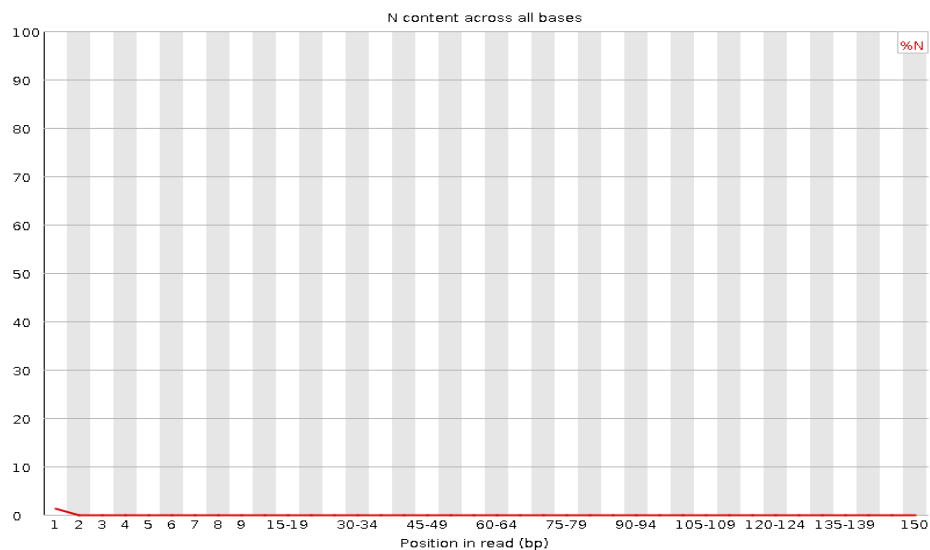


Figure B.28 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

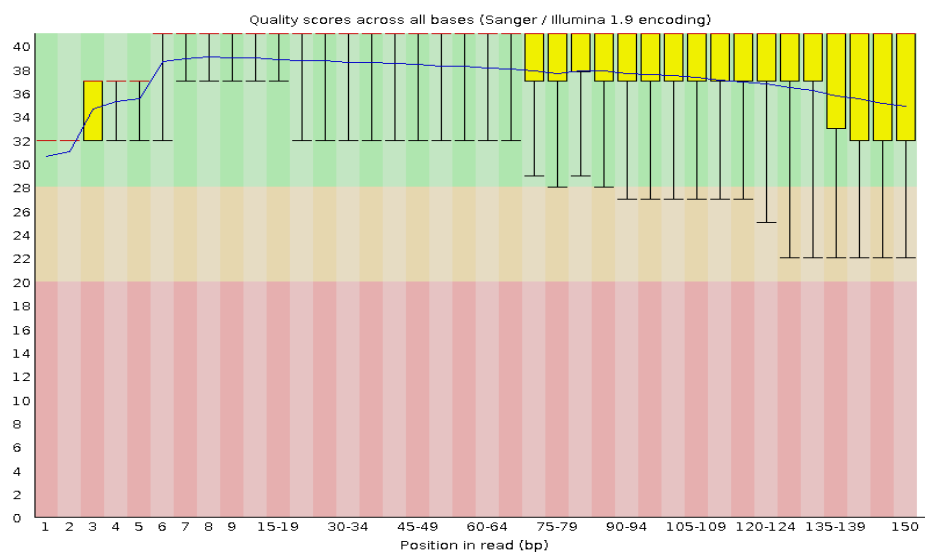


Figure B.29 Quality scores across the positions of the 150 bp reads for sample A2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

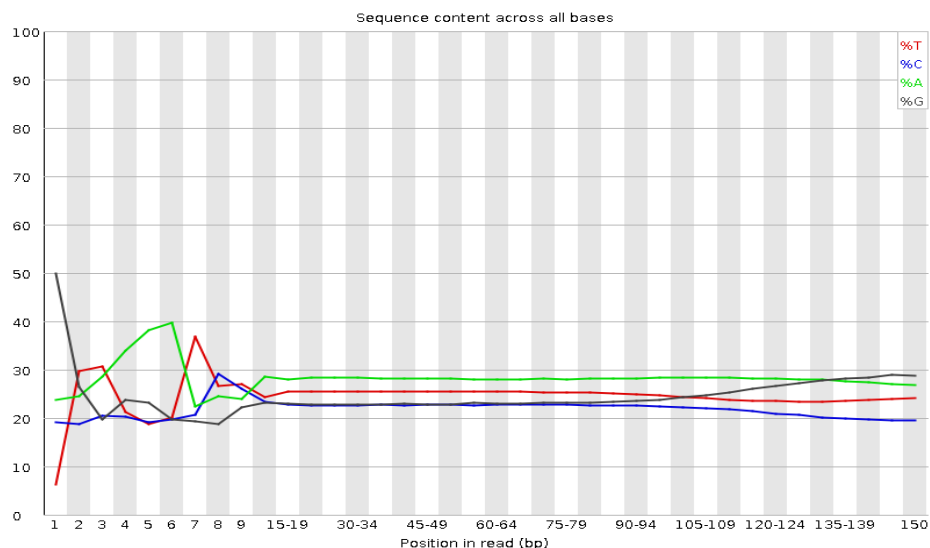


Figure B.30 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

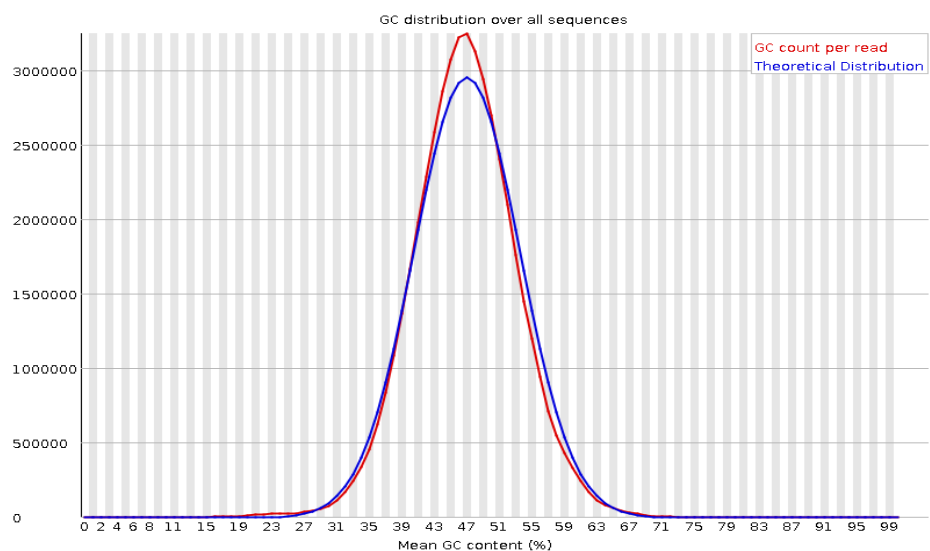


Figure B.31 The per sequence GC content for sample A2-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

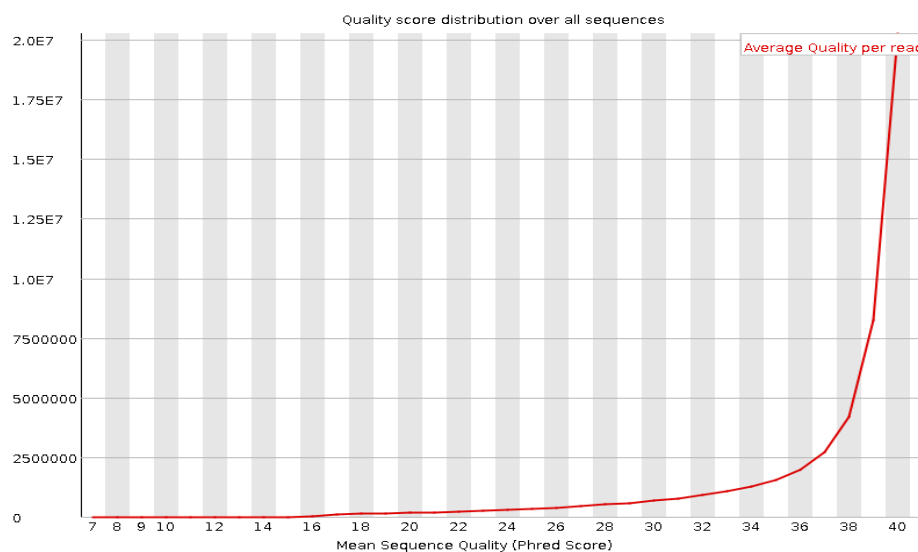


Figure B.32 The per sequence quality score for sample A2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.33 The distribution of the sequence lengths for sample A2-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample A3-R1

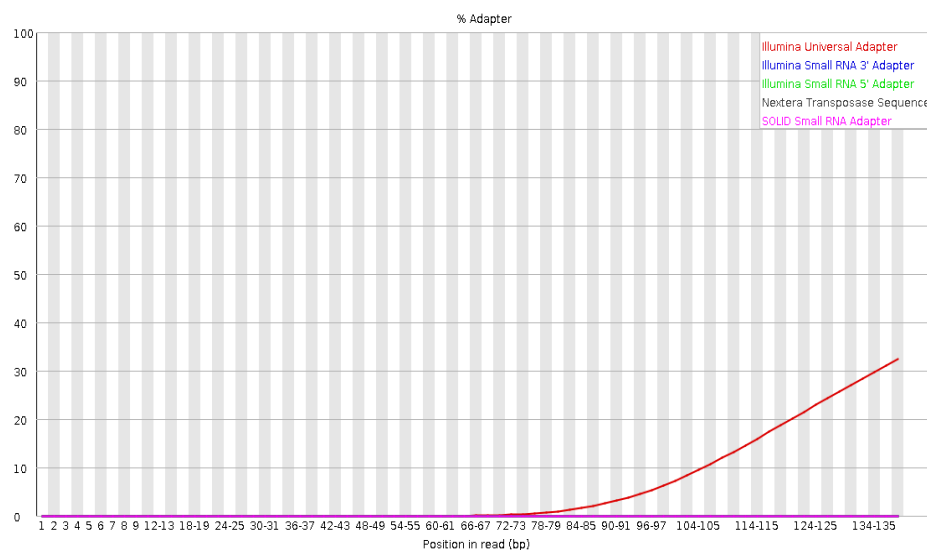


Figure B.34 This figure indicates the presence of adapter sequence contamination in sample A3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

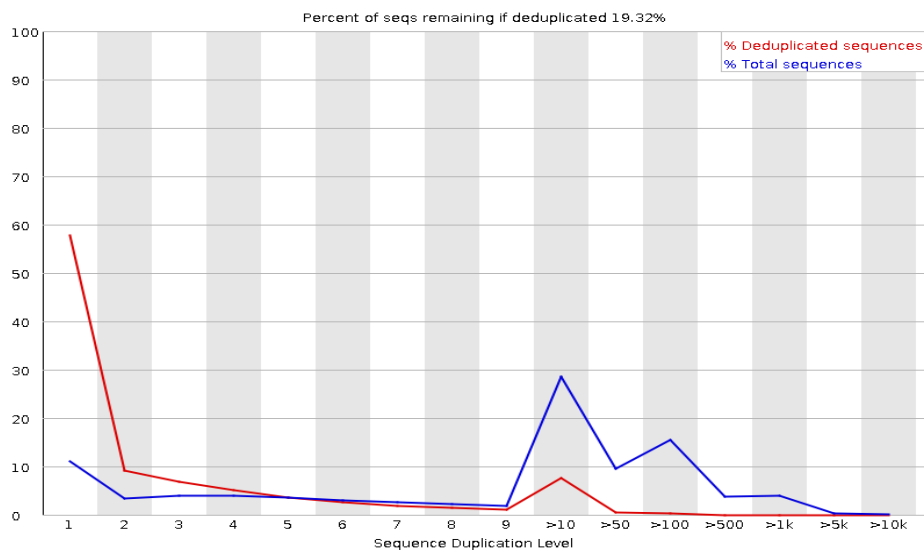


Figure B.35 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A3-R1. Here, 19.32% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

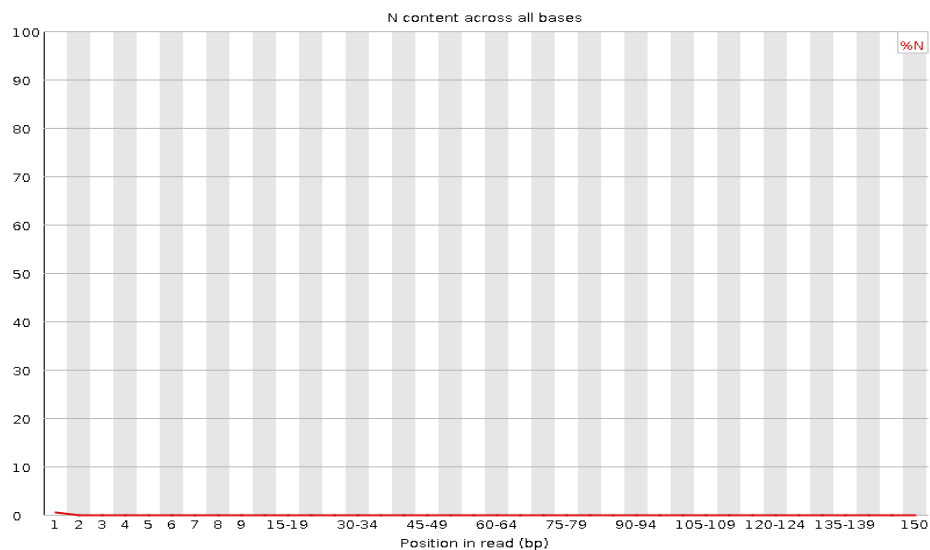


Figure B.36 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

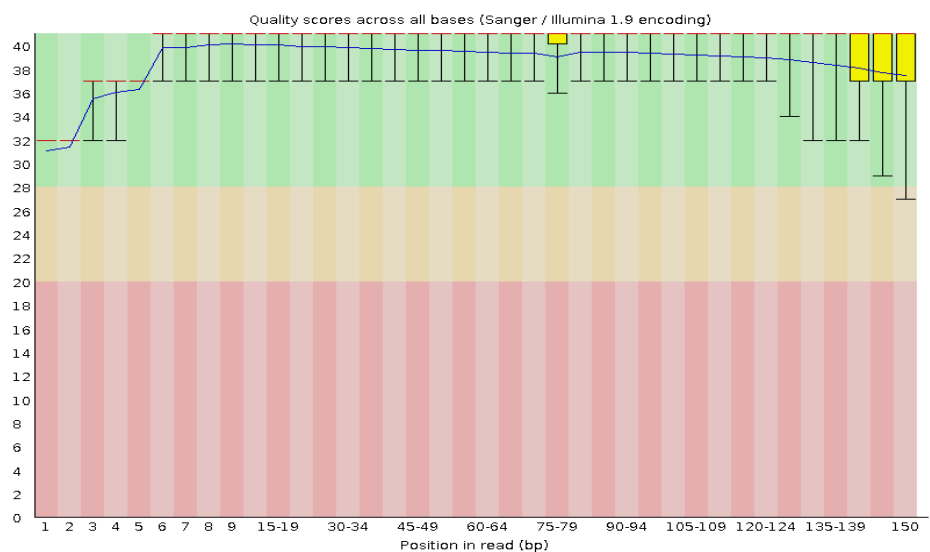


Figure B.37 Quality scores across the positions of the 150 bp reads for sample A3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

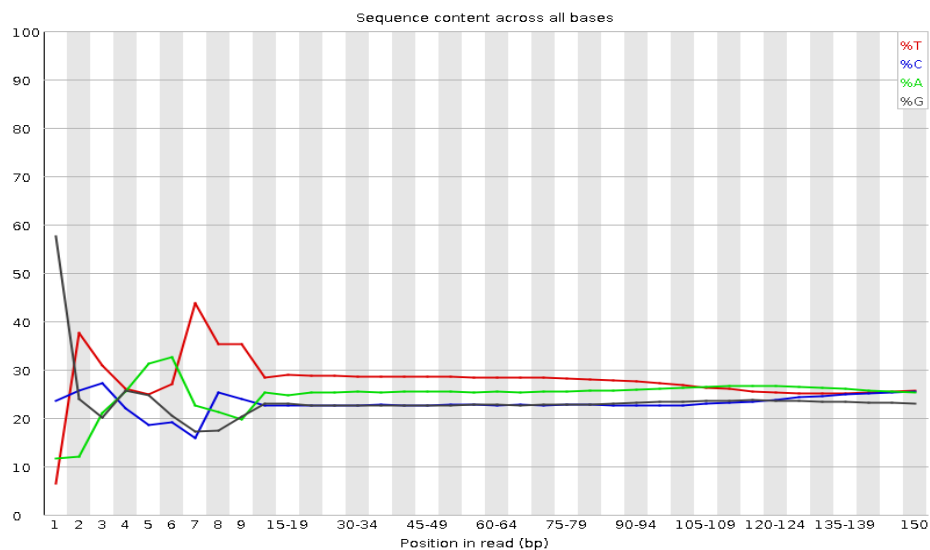


Figure B.38 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

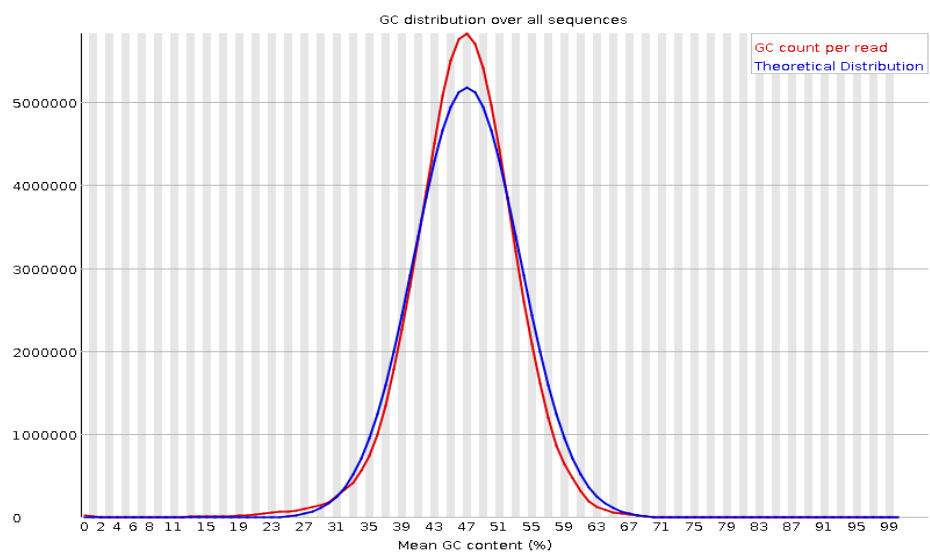


Figure B.39 The per sequence GC content for sample A3-R1. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

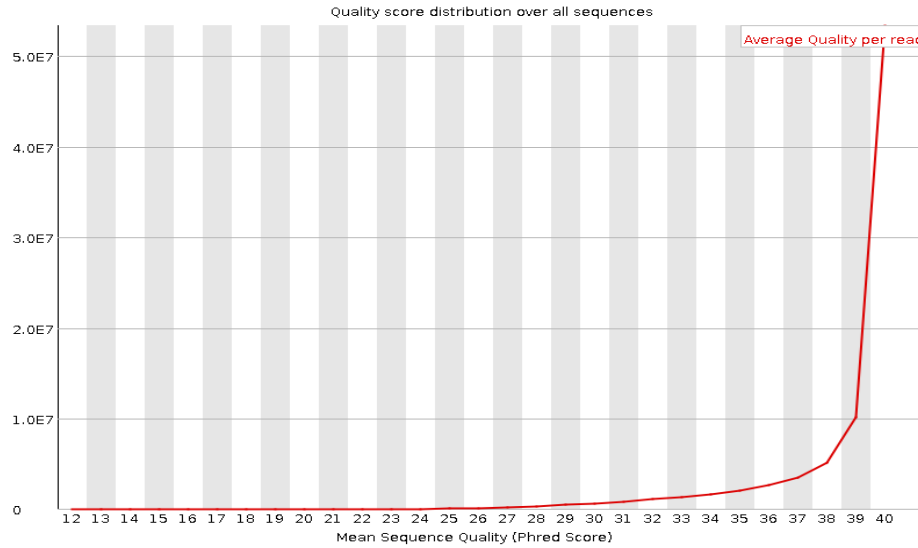


Figure B.40 The per sequence quality score for sample A3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .

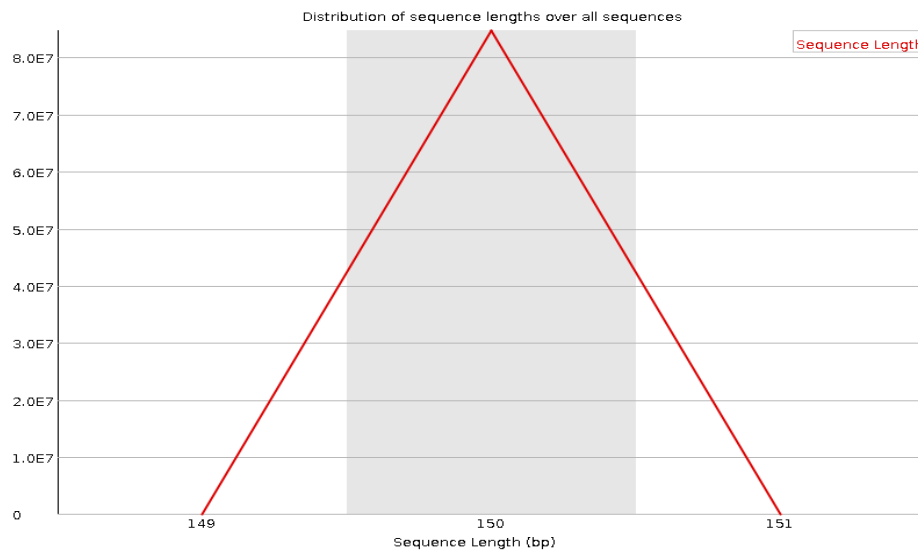


Figure B.41 The distribution of the sequence lengths for sample A3-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample A3-R2

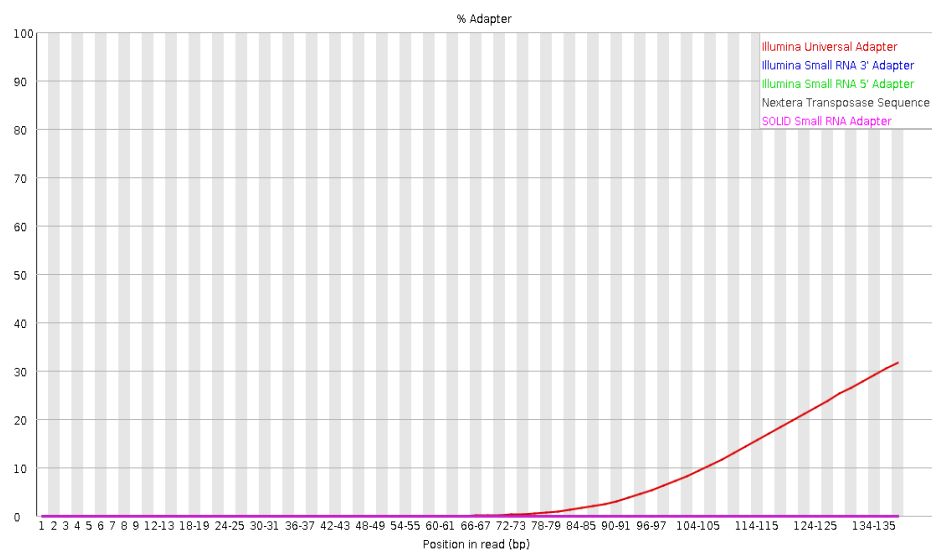


Figure B.42 This figure indicates the presence of adapter sequence contamination in sample A3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

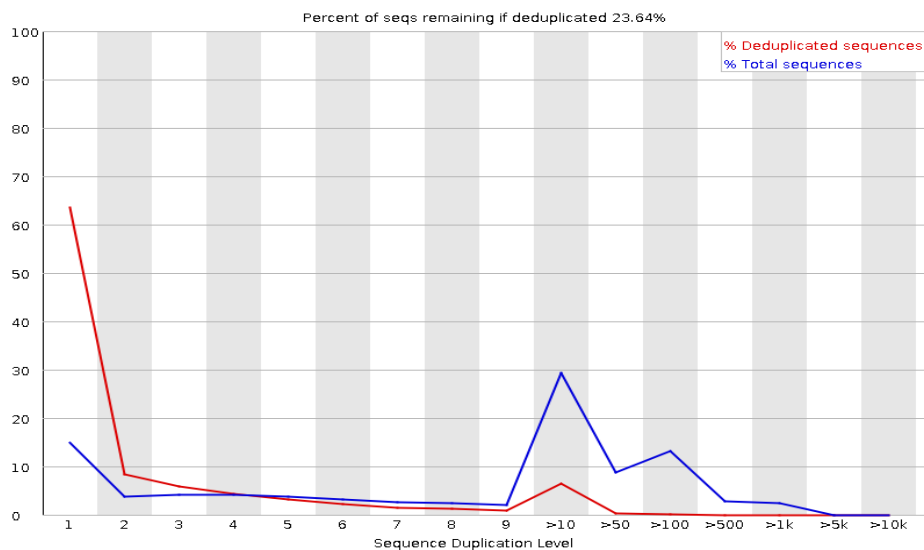


Figure B.43 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample A3-R2. Here, 23.64% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

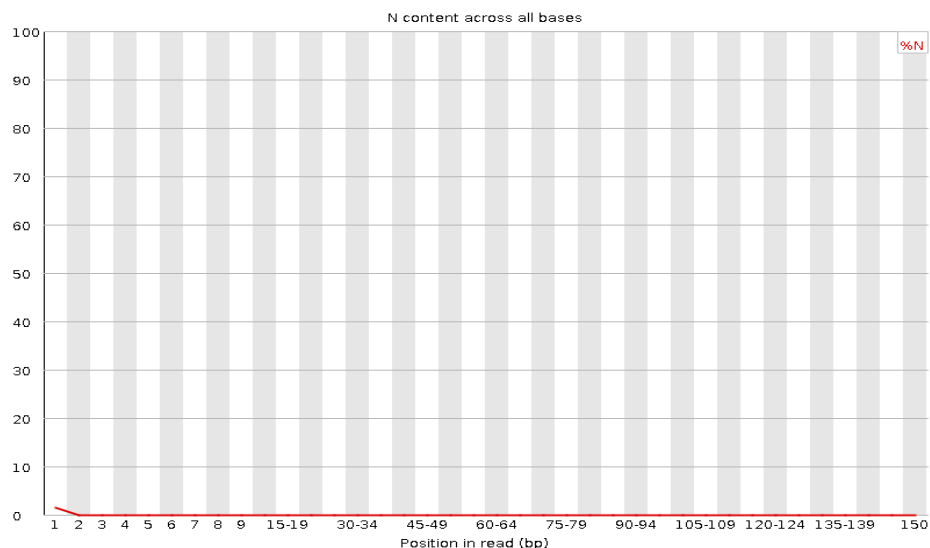


Figure B.44 The percentage of Ns and their positions across all bases in the 150 bp reads for sample A3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

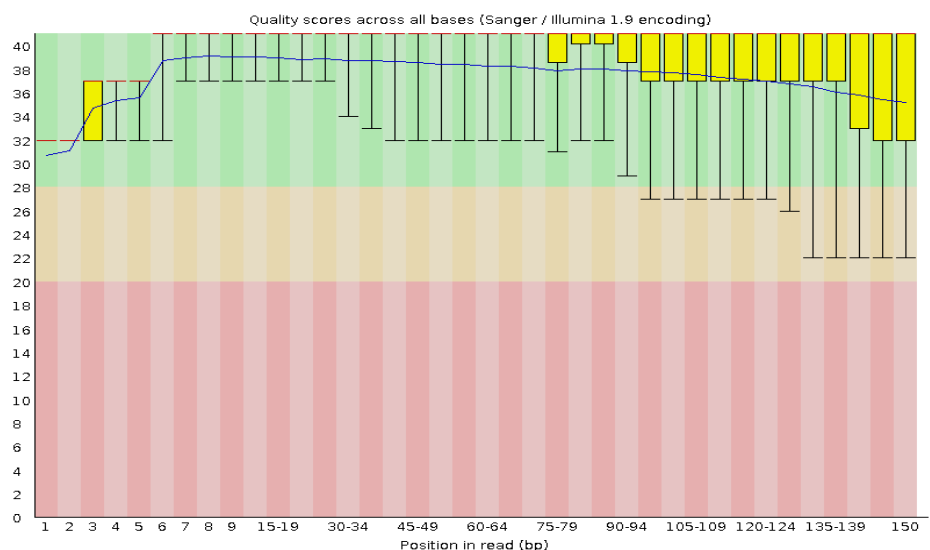


Figure B.45 Quality scores across the positions of the 150 bp reads for sample A3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

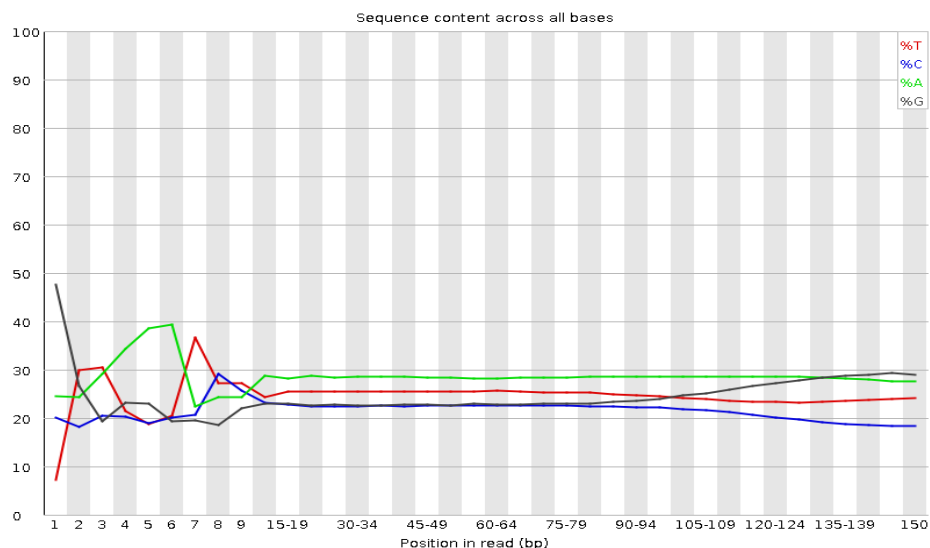


Figure B.46 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample A3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

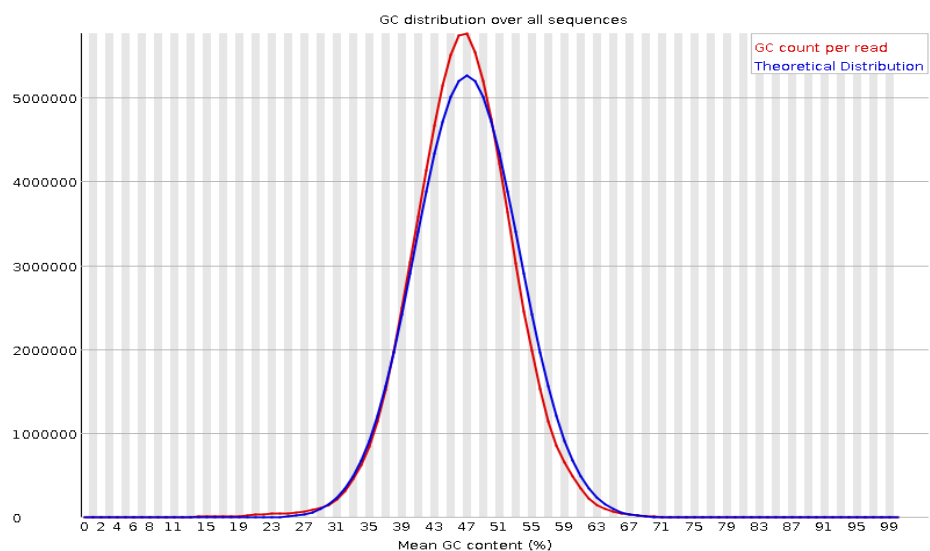


Figure B.47 The per sequence GC content for sample A3-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads.

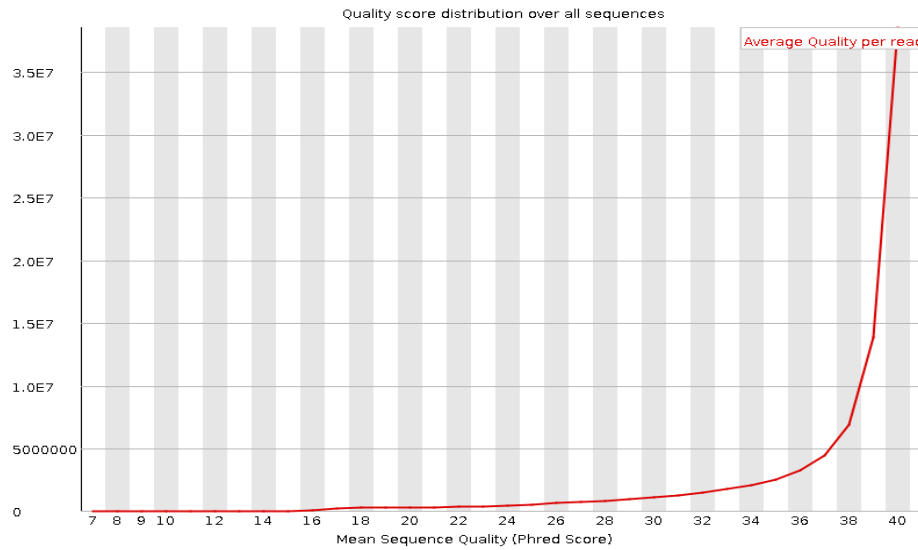


Figure B.48 The per sequence quality score for sample A3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.

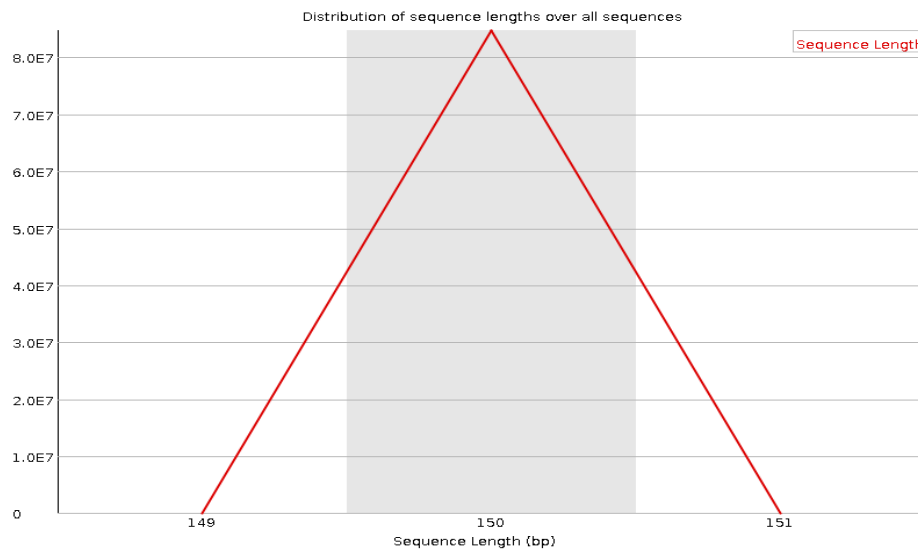


Figure B.49 The distribution of the sequence lengths for sample A3-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B1-R1

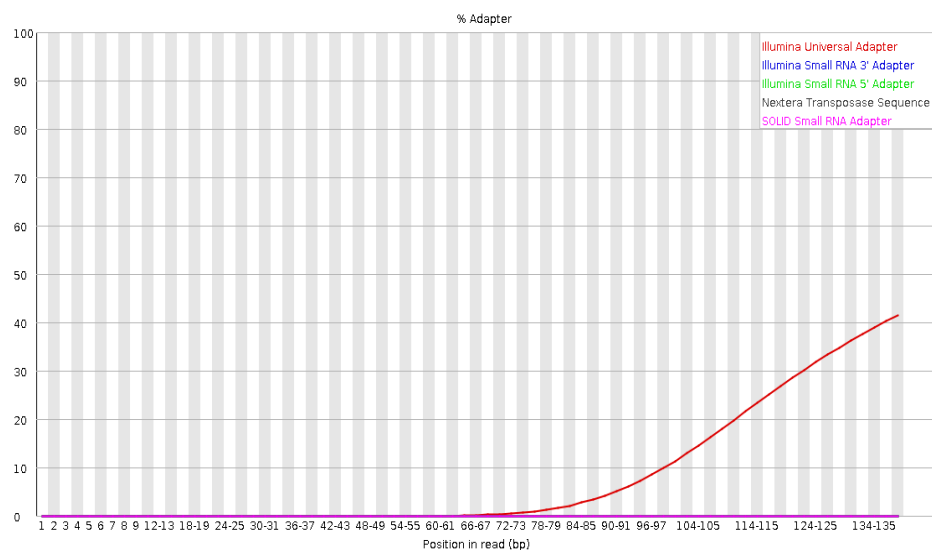


Figure B.50 This figure indicates the presence of adapter sequence contamination in sample B1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

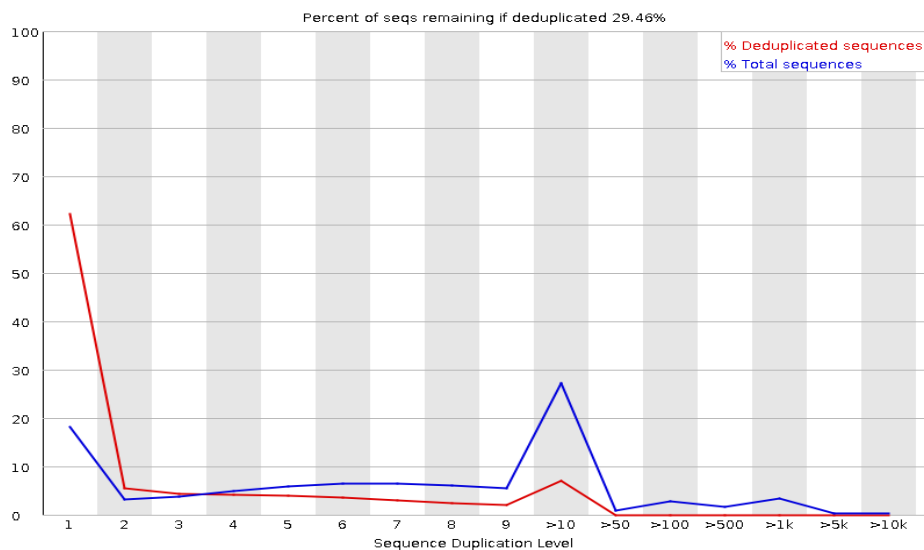


Figure B.51 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B1-R1. Here, 29.46% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

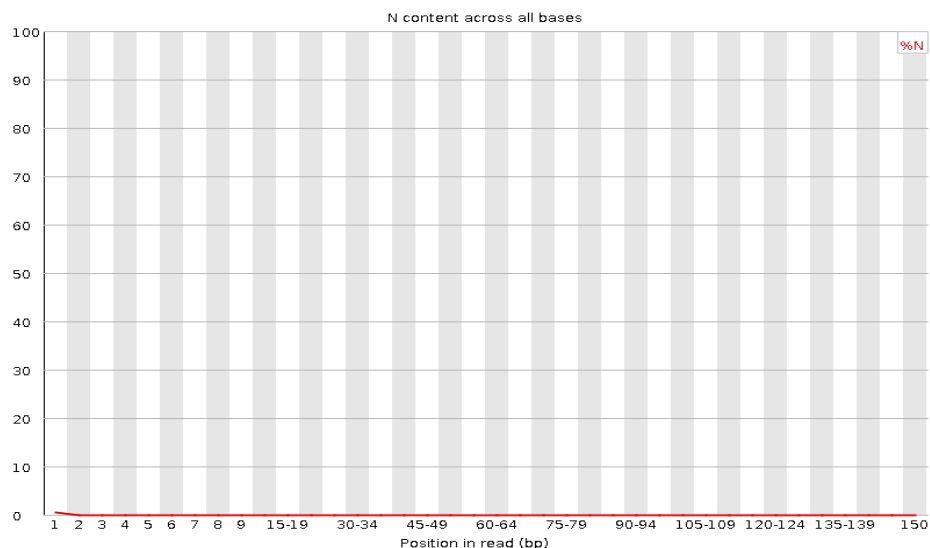


Figure B.52 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

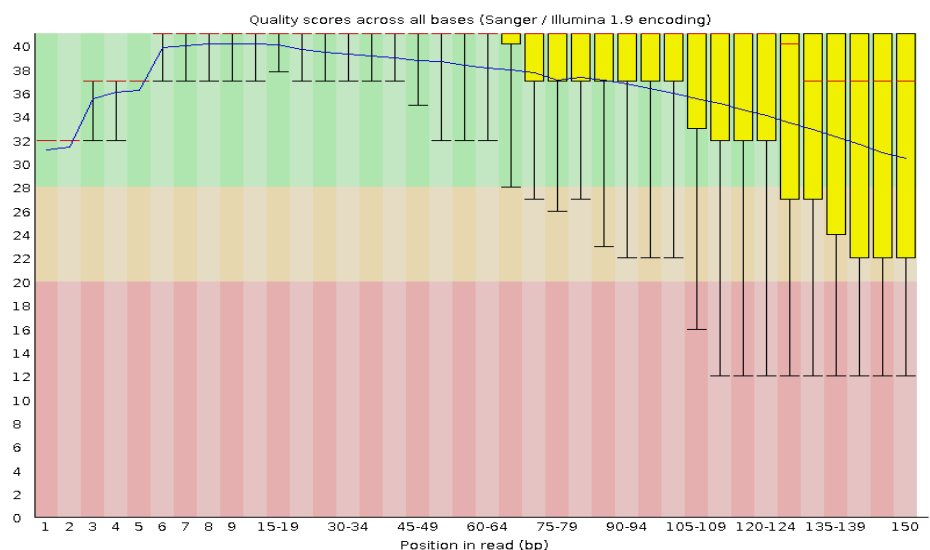


Figure B.53 Quality scores across the positions of the 150 bp reads for sample B1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

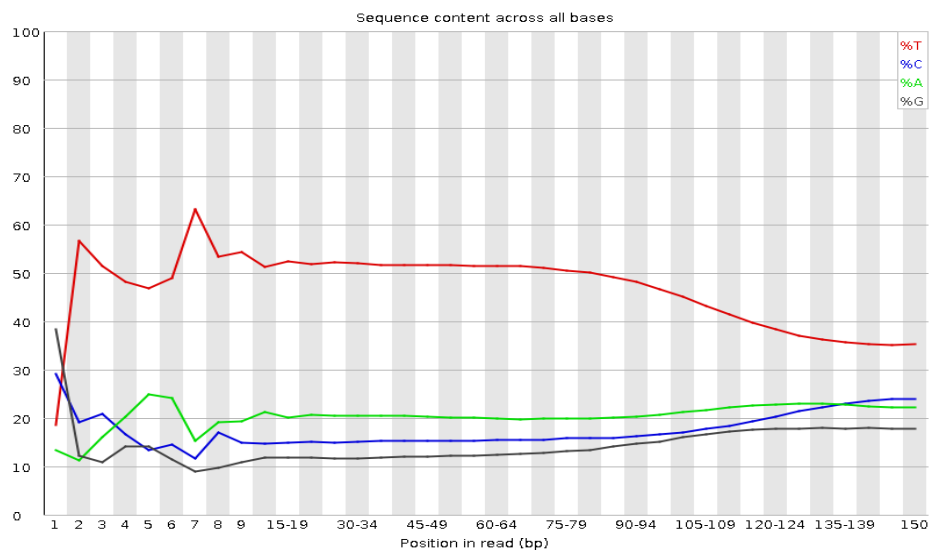


Figure B.54 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

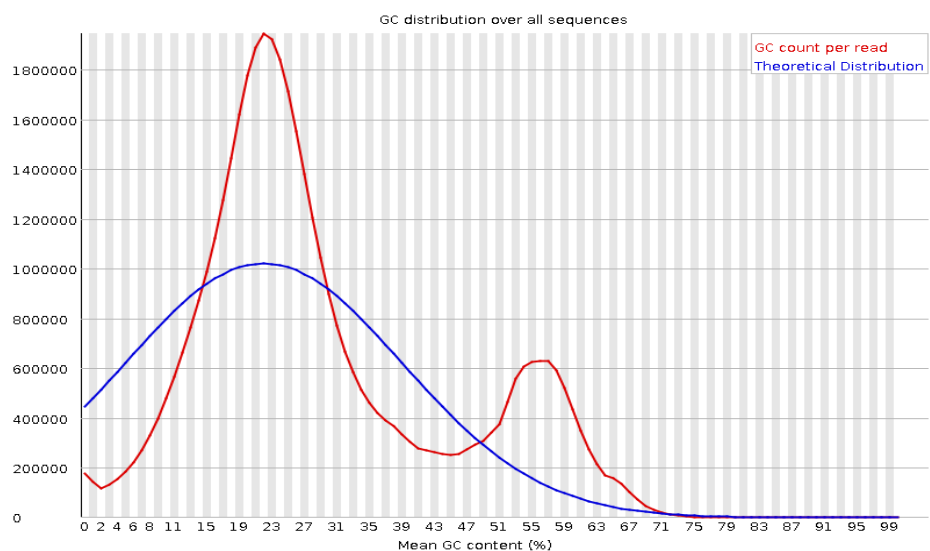


Figure B.55 The per sequence GC content for sample B1-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

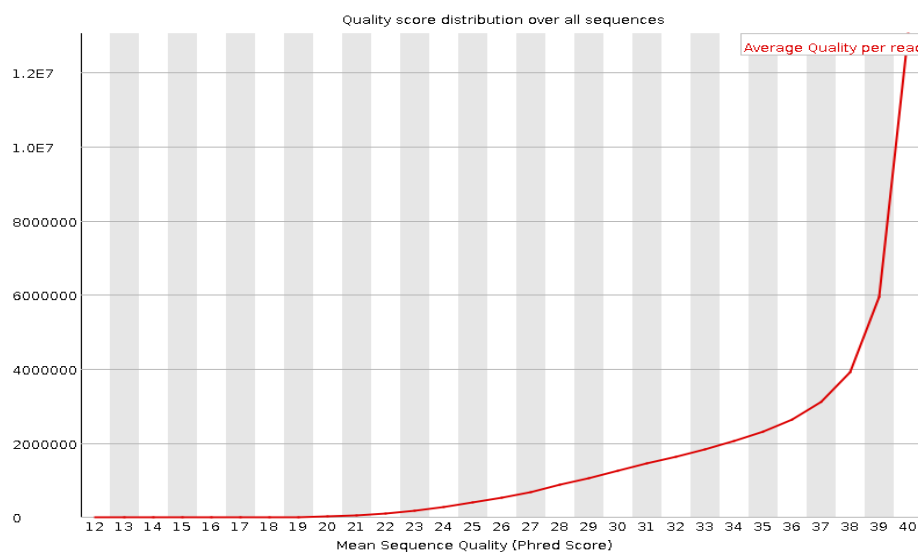


Figure B.56 The per sequence quality score for sample B1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.57 The distribution of the sequence lengths for sample B1-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B1-R2

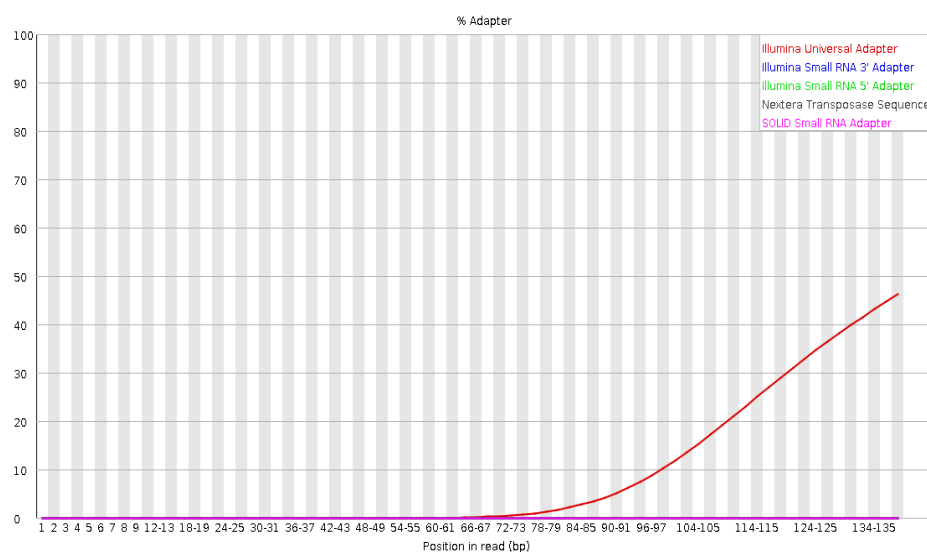


Figure B.58 This figure indicates the presence of adapter sequence contamination in sample B1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

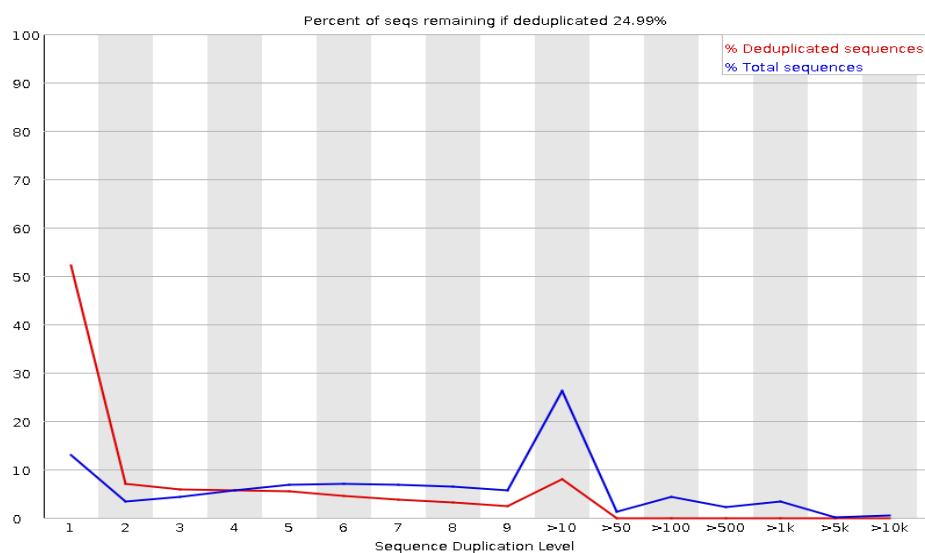


Figure B.59 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B1-R2. Here, 24.99% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

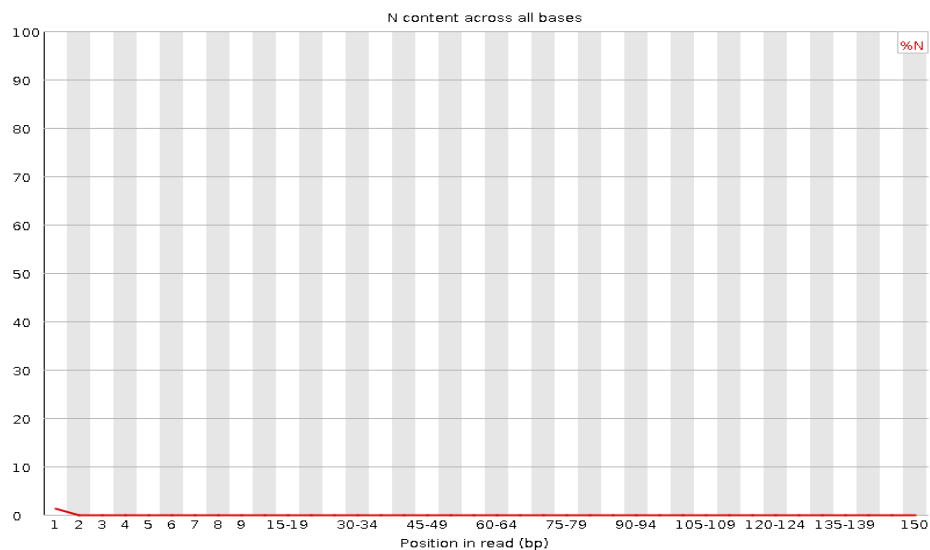


Figure B.60 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

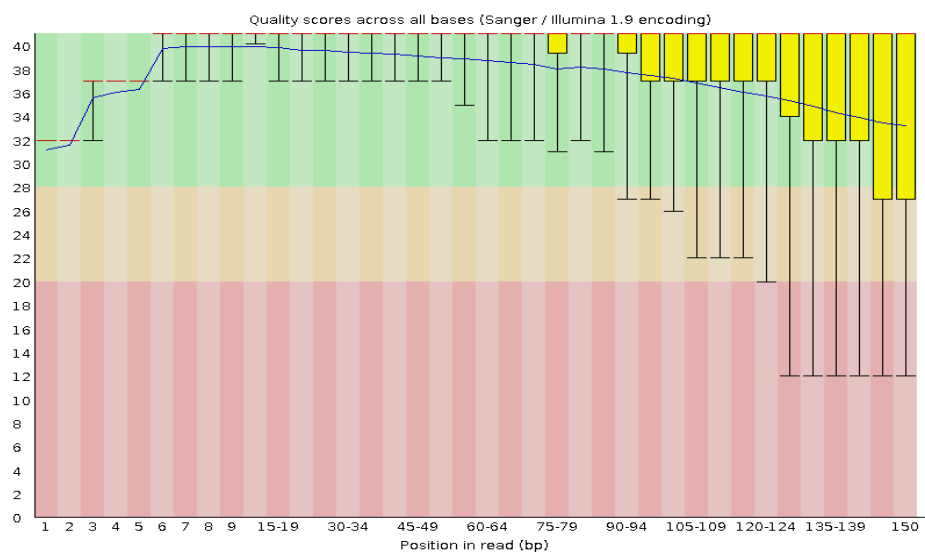


Figure B.61 Quality scores across the positions of the 150 bp reads for sample B1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

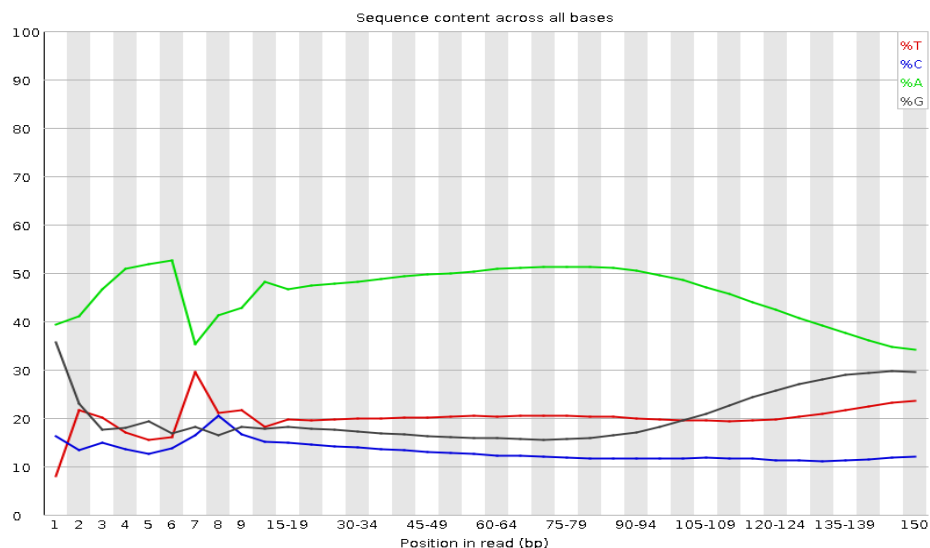


Figure B.62 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

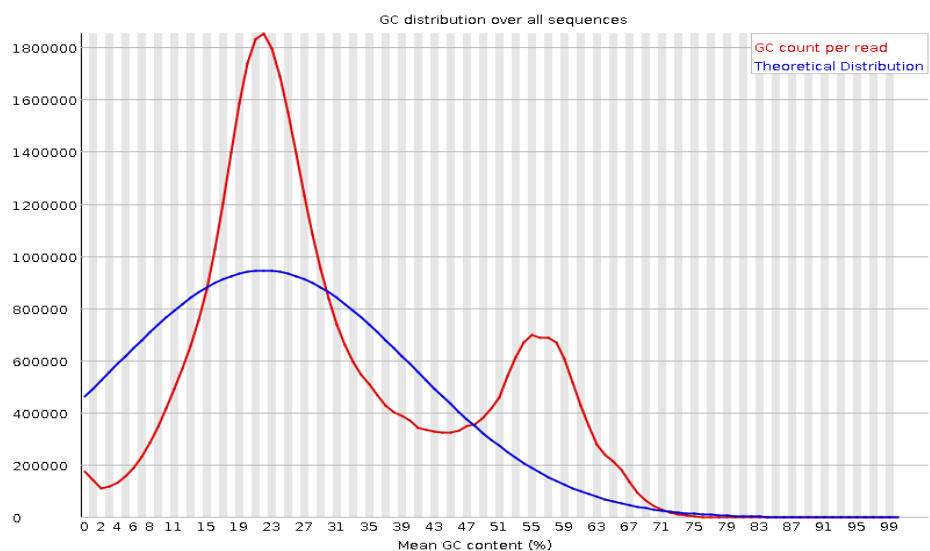


Figure B.63 The per sequence GC content for sample B1-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

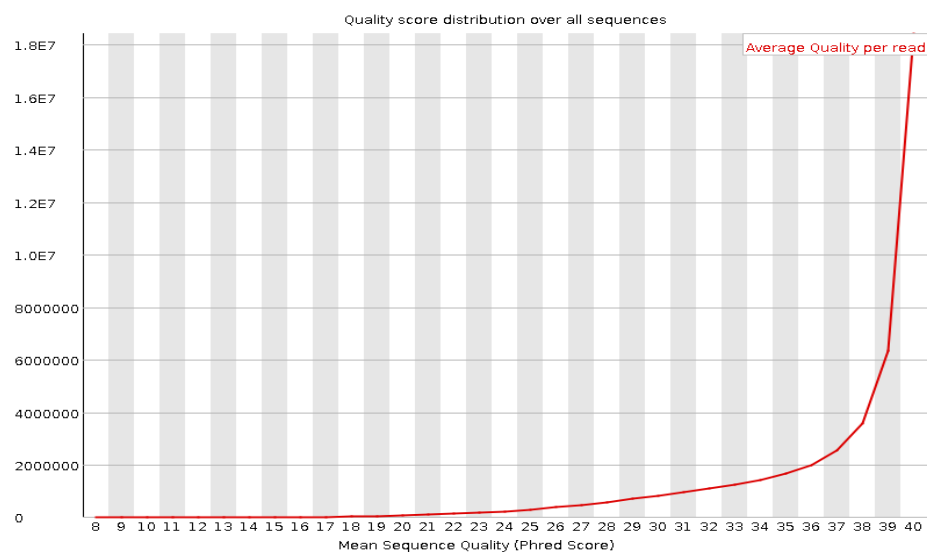


Figure B.64 The per sequence quality score for sample B1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.65 The distribution of the sequence lengths for sample B1-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B2-R1

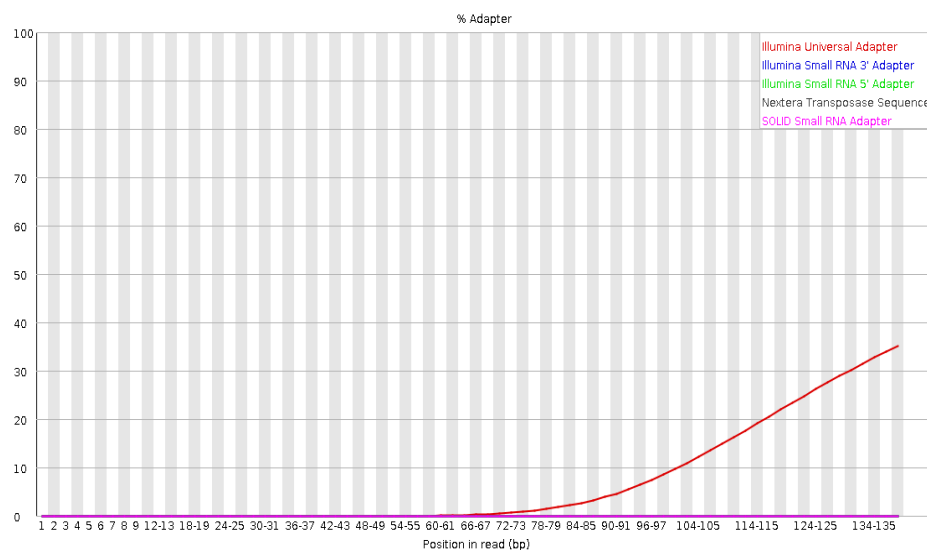


Figure B.66 This figure indicates the presence of adapter sequence contamination in sample B2-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

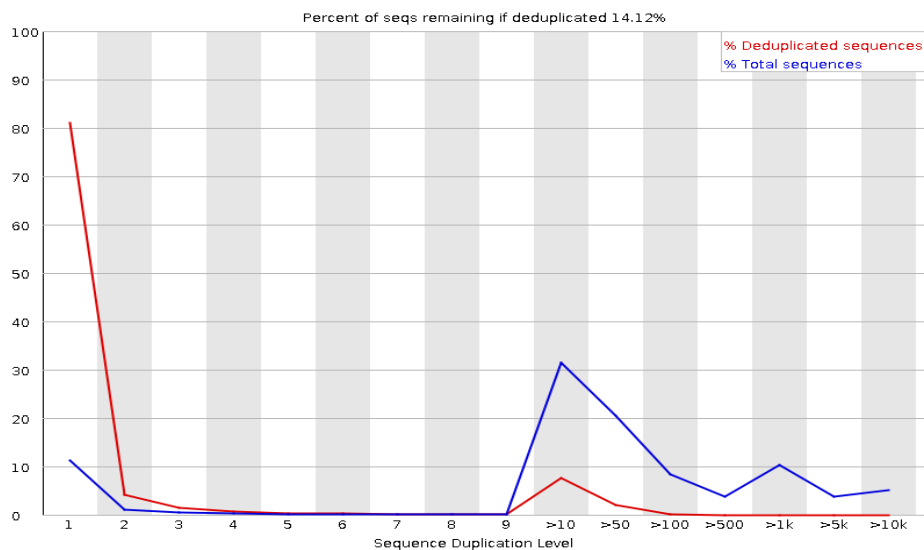


Figure B.67 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B2-R1. Here, 14.12% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

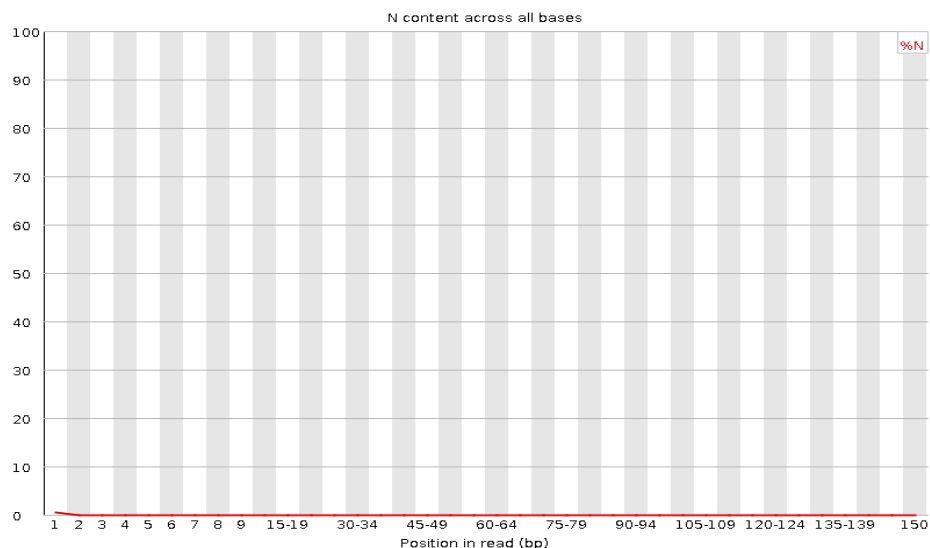


Figure B.68 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B2-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

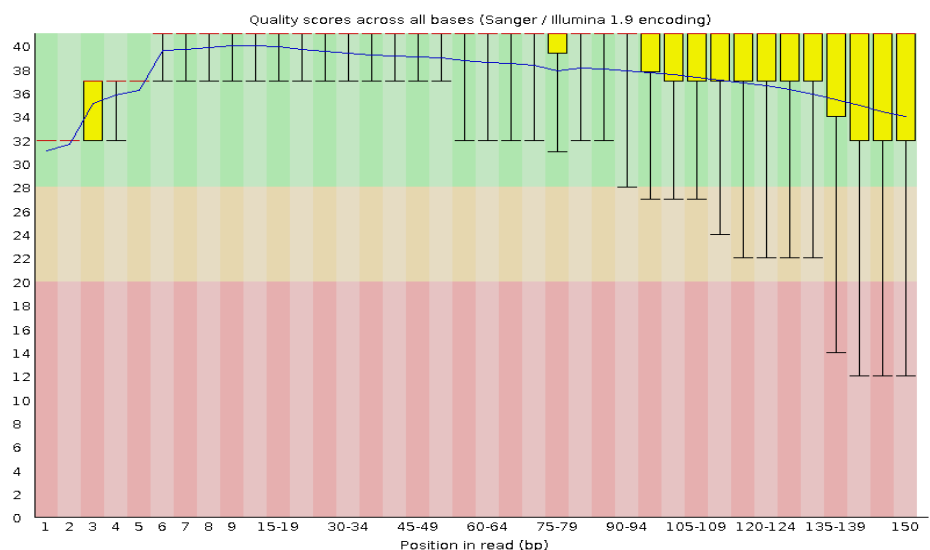


Figure B.69 Quality scores across the positions of the 150 bp reads. The x- and y-axes represent the quality scores and position within the read for sample B2-R1. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

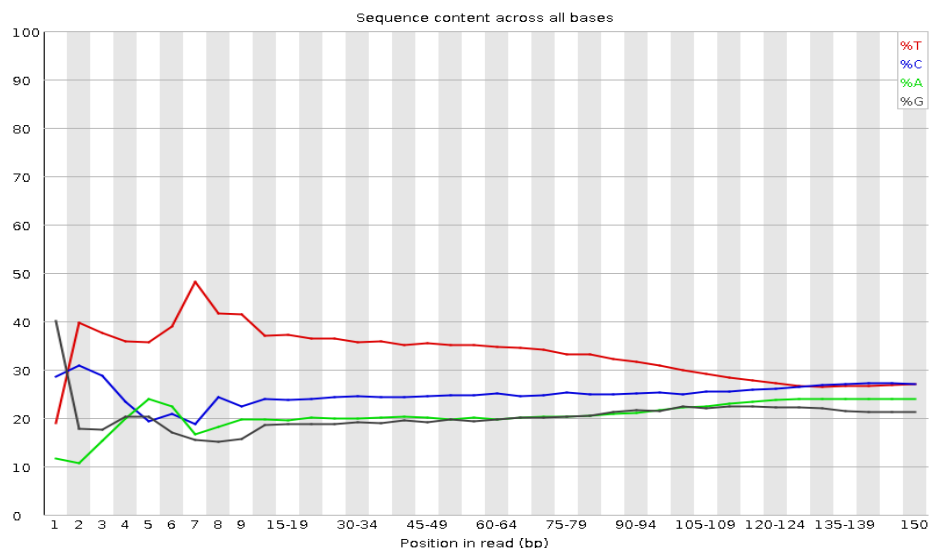


Figure B.70 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B2-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

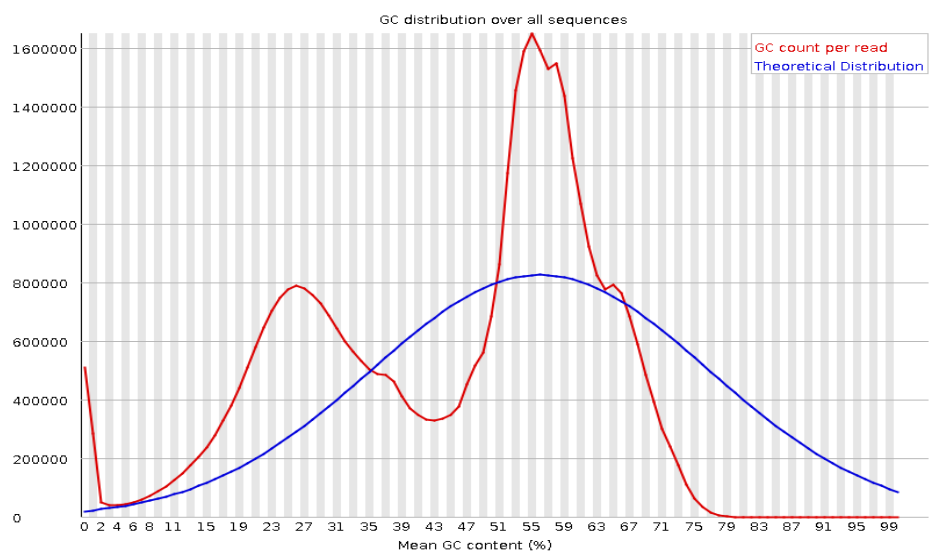


Figure B.71 The per sequence GC content for sample B2-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

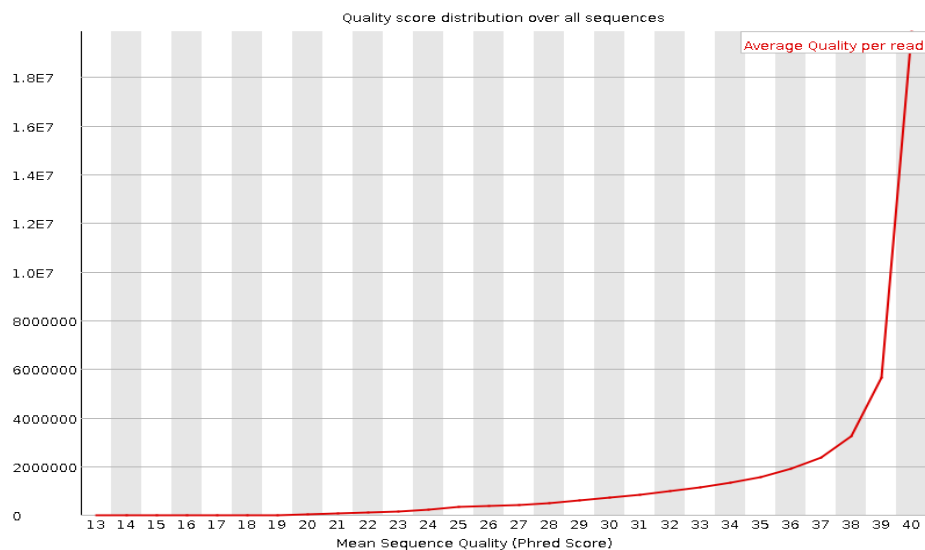


Figure B.72 The per sequence quality score for sample B2-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.73 The distribution of the sequence lengths for sample B2-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B2-R2

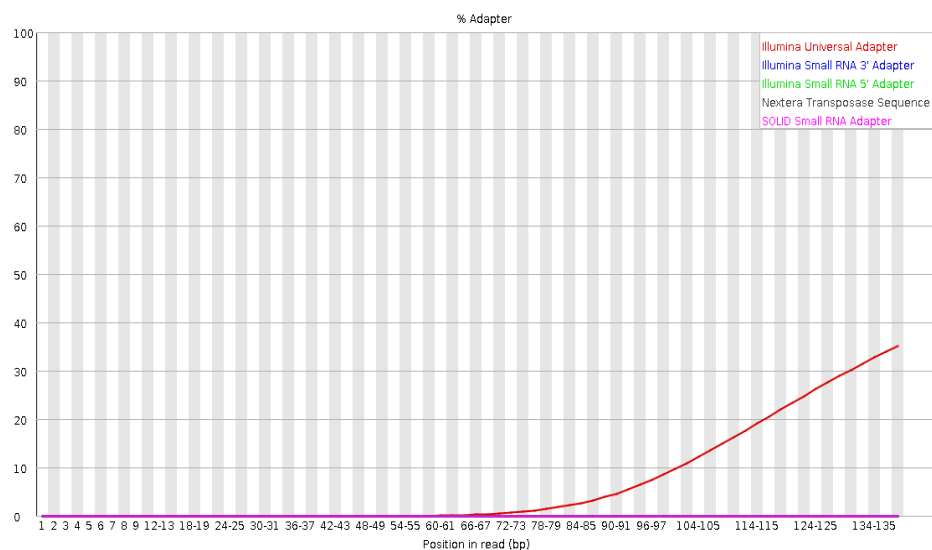


Figure B.74 This figure indicates the presence of adapter sequence contamination in sample B2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

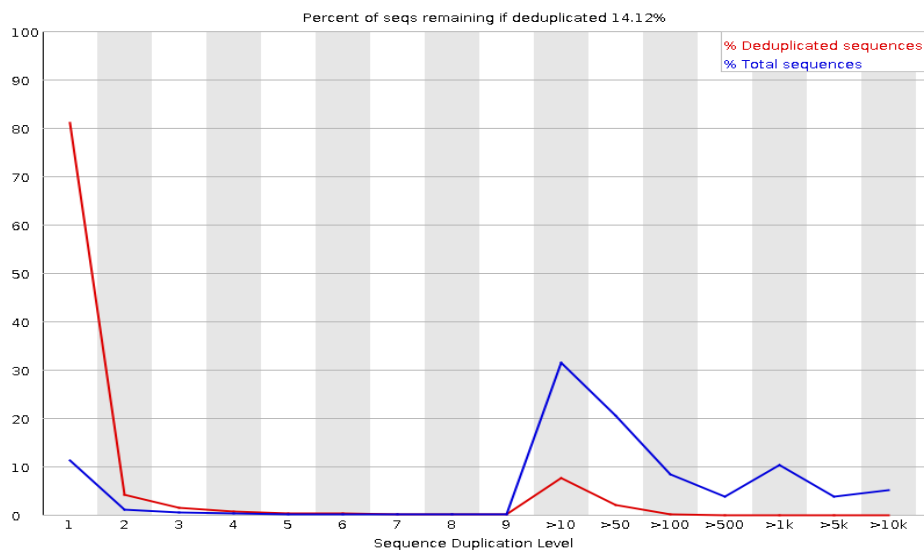


Figure B.75 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B2-R2. Here, 14.12% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

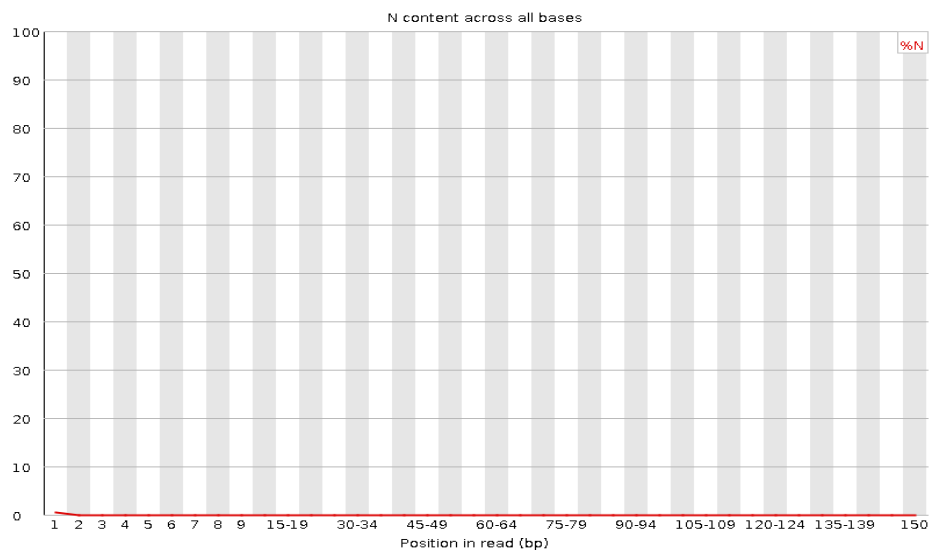


Figure B.76 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

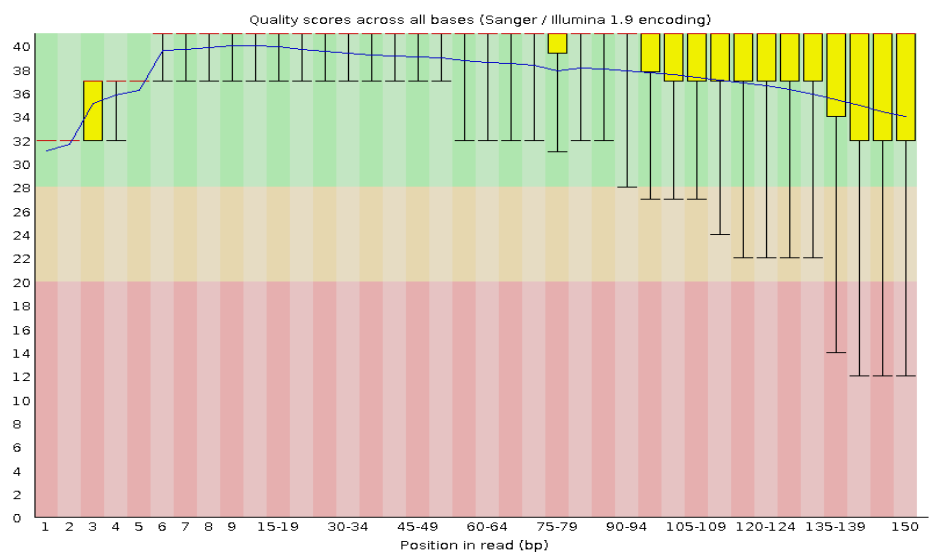


Figure B.77 Quality scores across the positions of the 150 bp reads for sample B2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

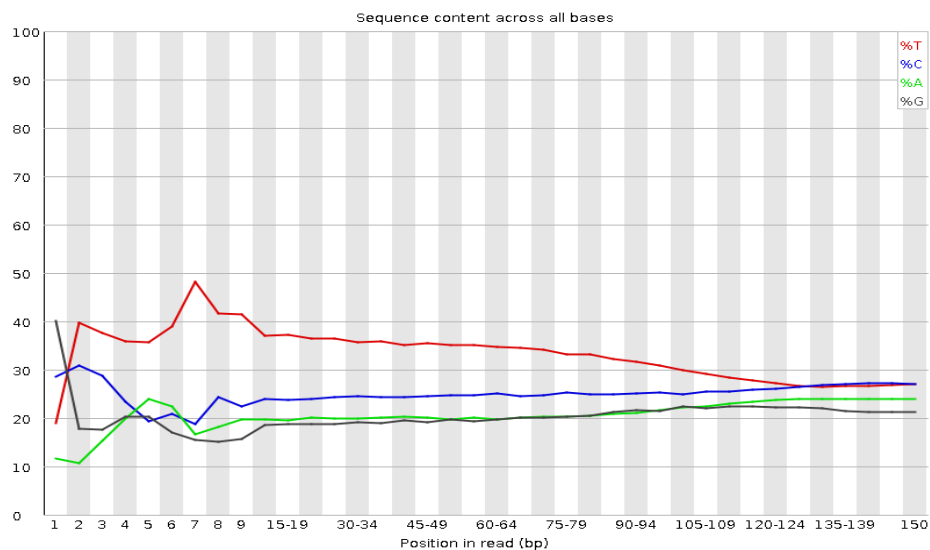


Figure B.78 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

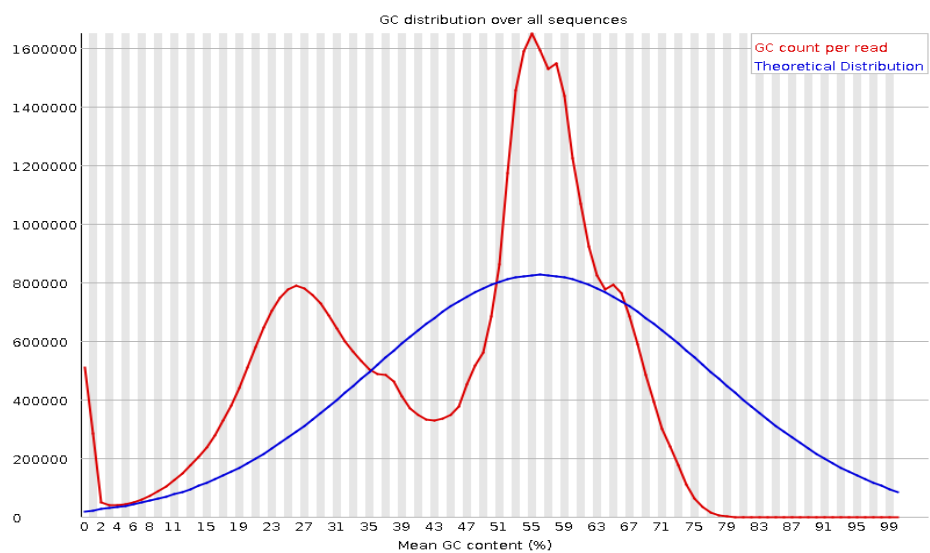


Figure B.79 The per sequence GC content for sample B2-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

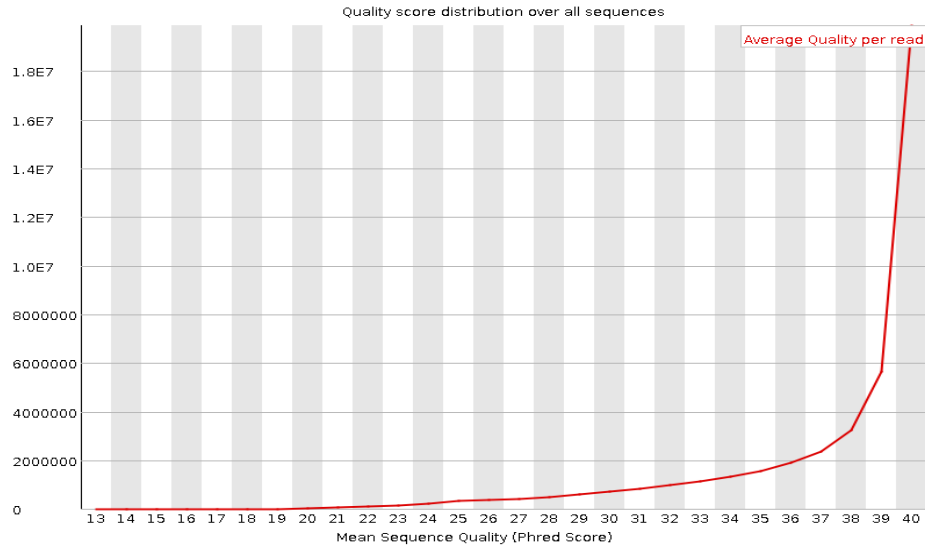


Figure B.80 The per sequence quality score for sample B2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.

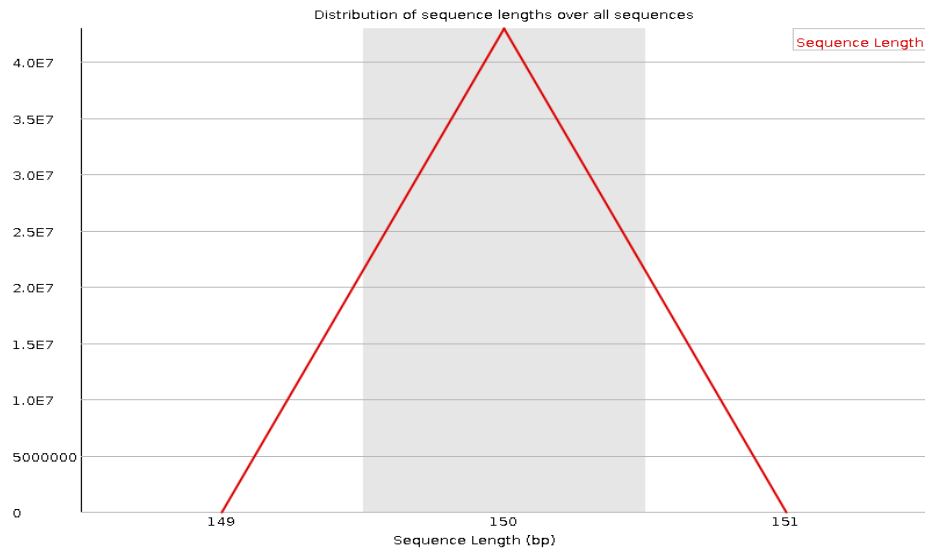


Figure B.81 The distribution of the sequence lengths for sample B2-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B3-R1

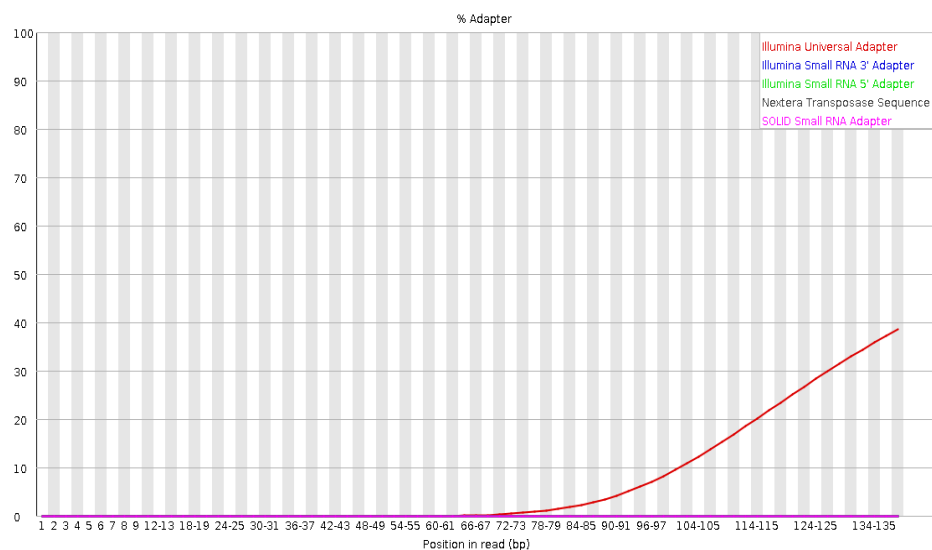


Figure B.82 This figure indicates the presence of adapter sequence contamination in sample B3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

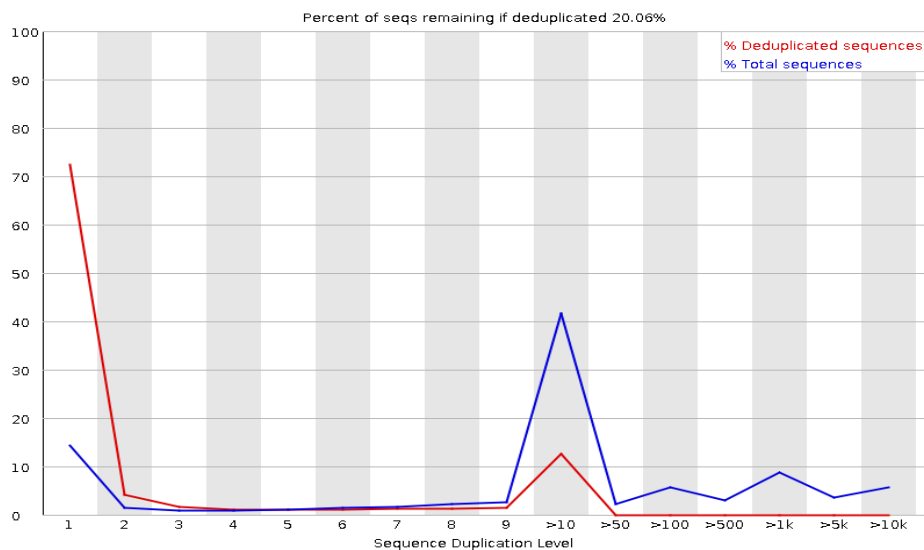


Figure B.83 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B3-R1. Here, 20.06% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

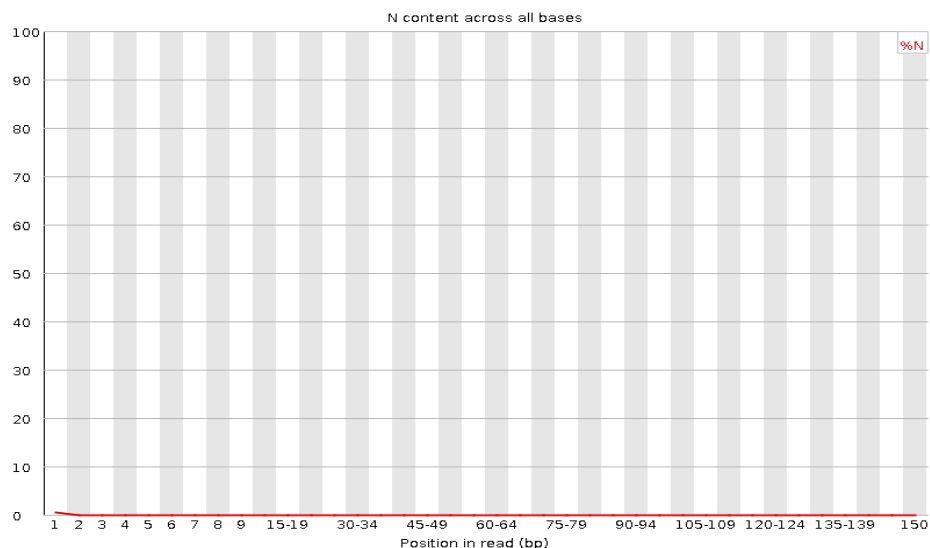


Figure B.84 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

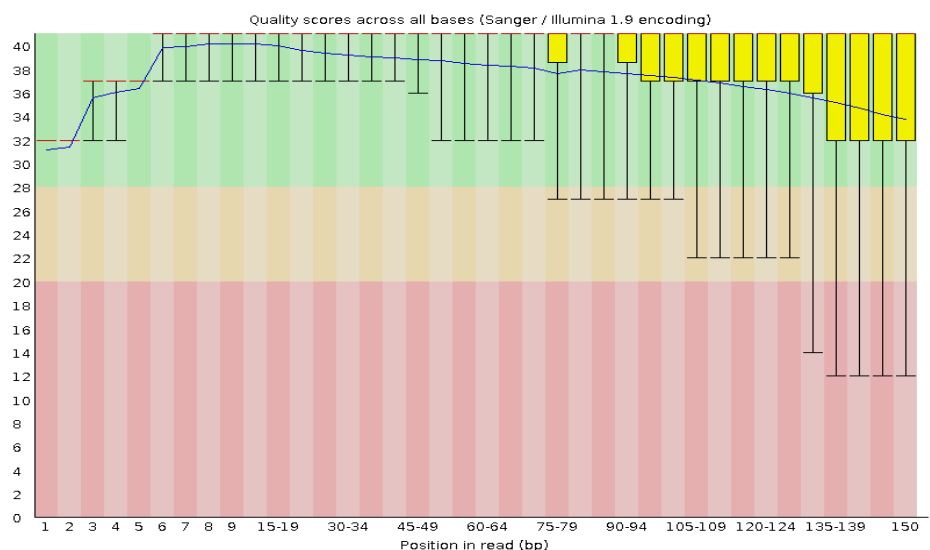


Figure B.85 Quality scores across the positions of the 150 bp reads for sample B3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

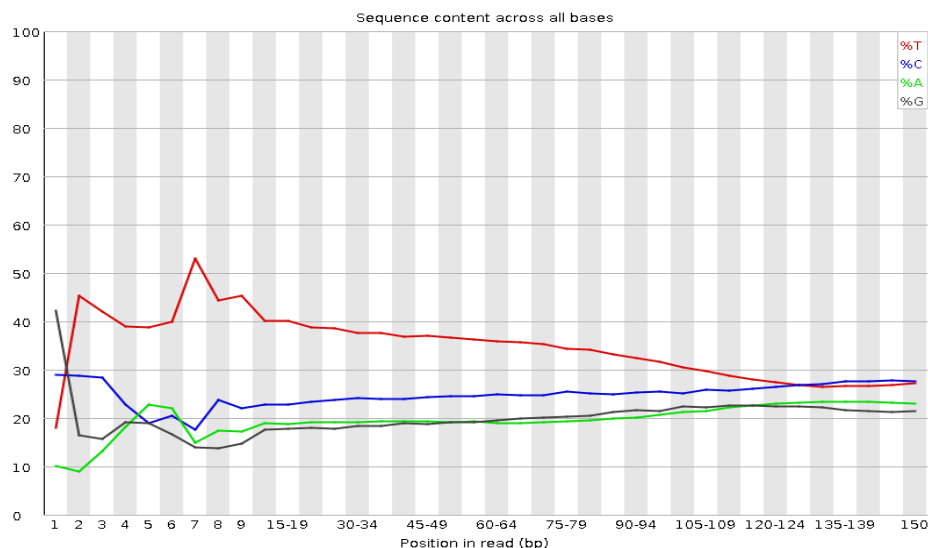


Figure B.86 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

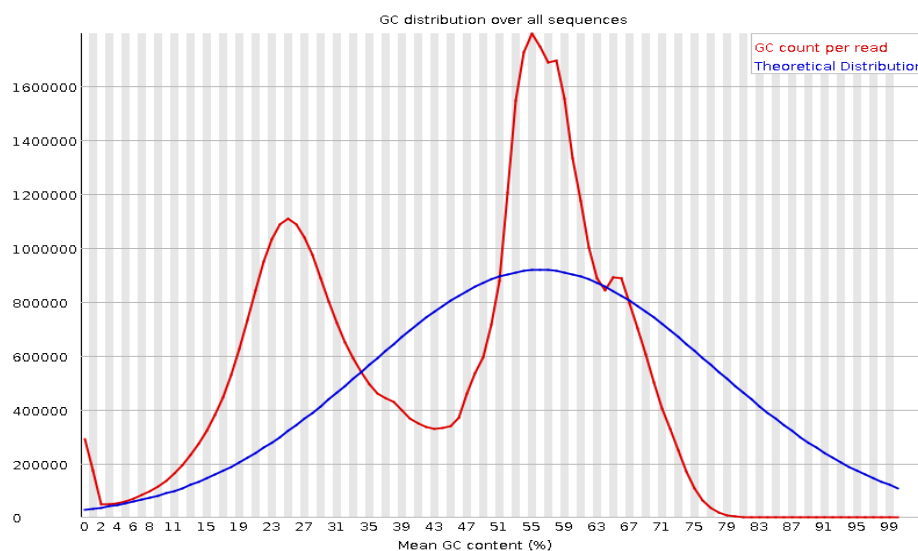


Figure B.87 The per sequence GC content for sample B3-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

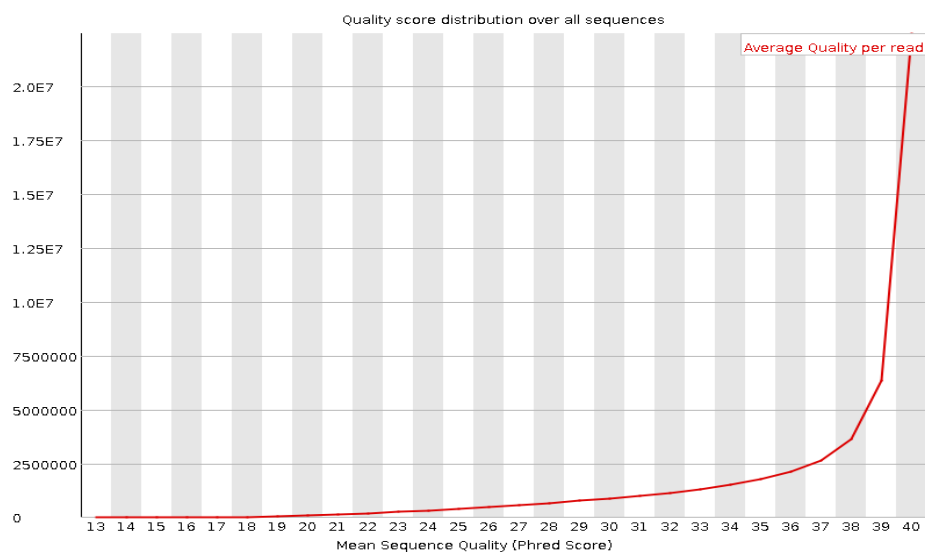


Figure B.88 The per sequence quality score for sample B3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.89 The distribution of the sequence lengths for sample B3-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample B3-R2

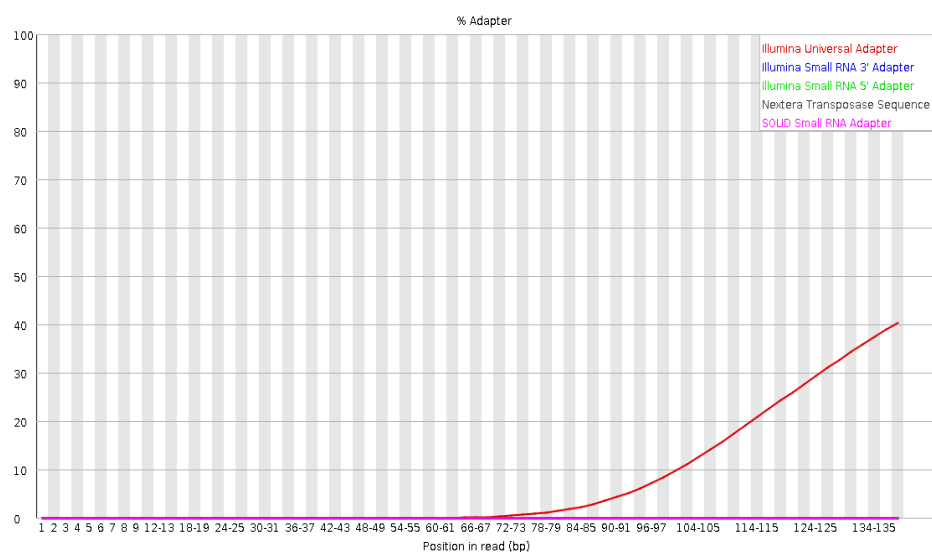


Figure B.90 This figure indicates the presence of adapter sequence contamination in sample B3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

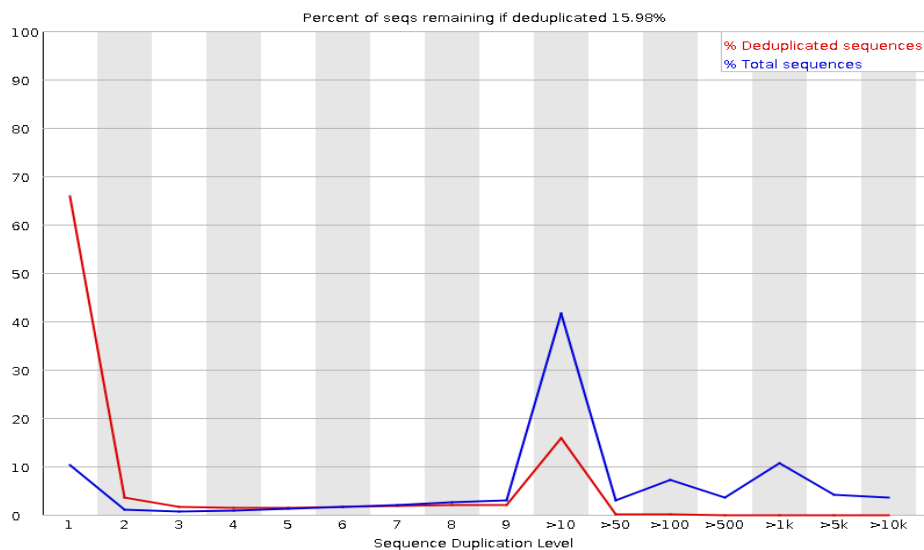


Figure B.90 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample B3-R2. Here, 15.98% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

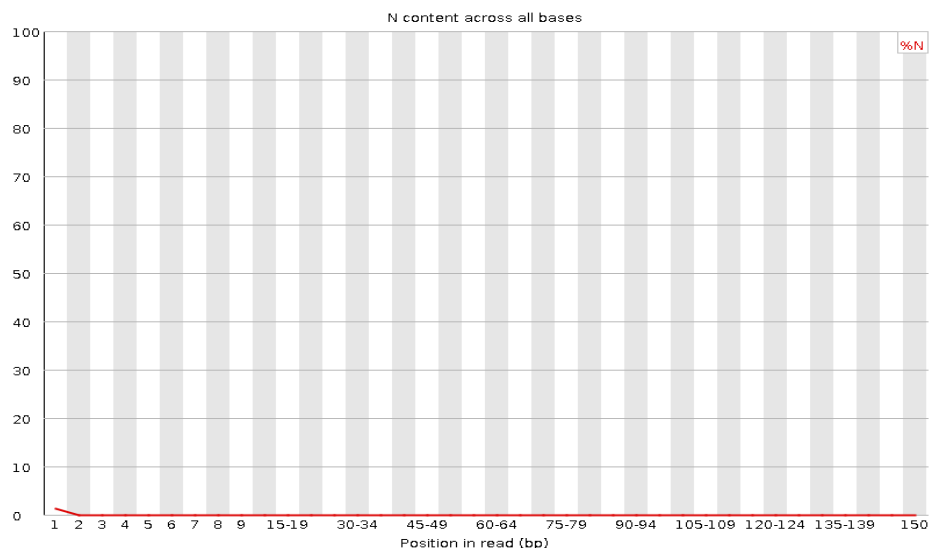


Figure B.91 The percentage of Ns and their positions across all bases in the 150 bp reads for sample B3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

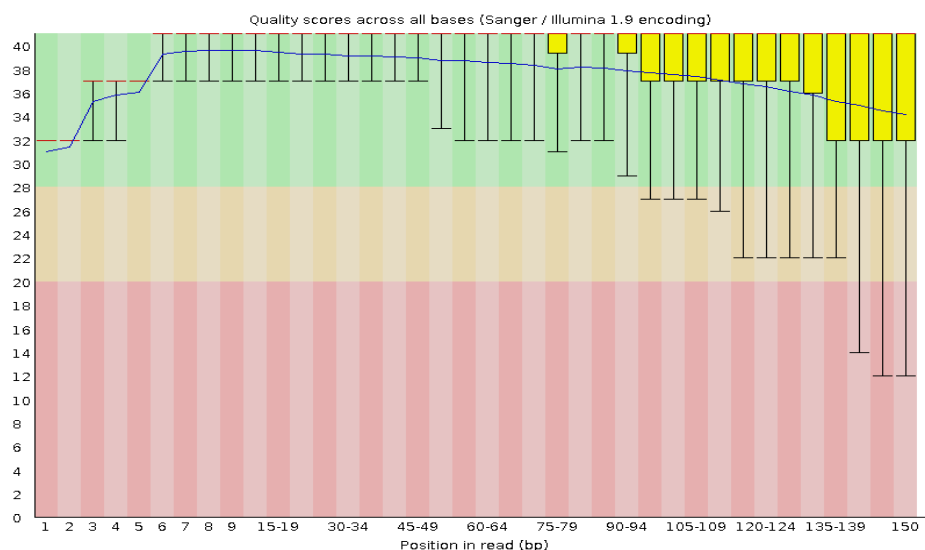


Figure 3 Quality scores across the positions of the 150 bp reads for sample B3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

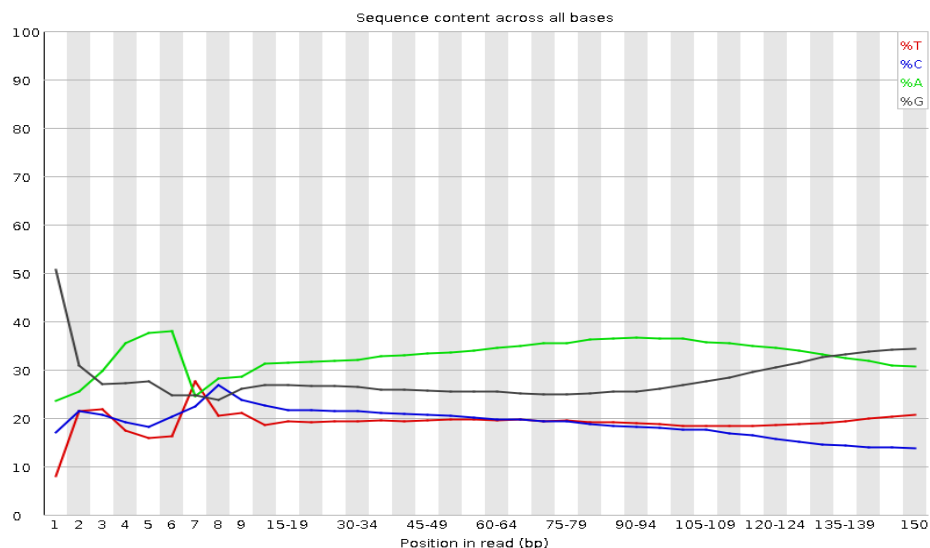


Figure B.93 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample B3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

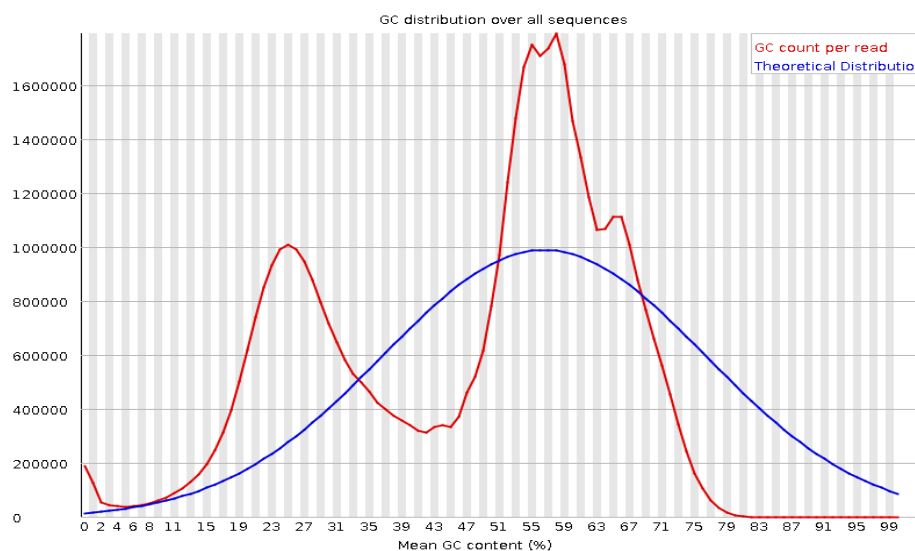


Figure B.94 The per sequence GC content for sample B3-R2. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The two broad, red peaks that shift away from the mean %GC = 47 indicates contamination.

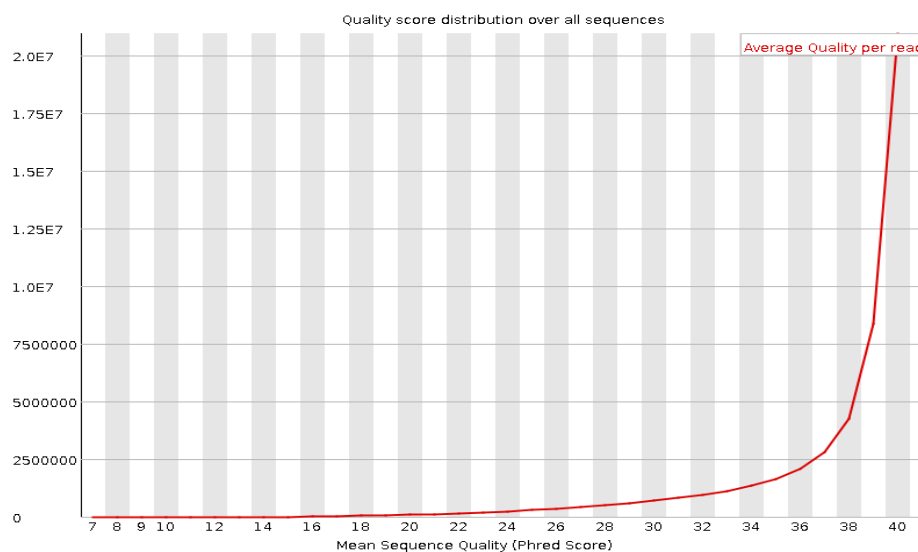


Figure B.96 The per sequence quality score for sample B3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.



Figure B.97 The distribution of the sequence lengths for sample B3-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample C1-R1

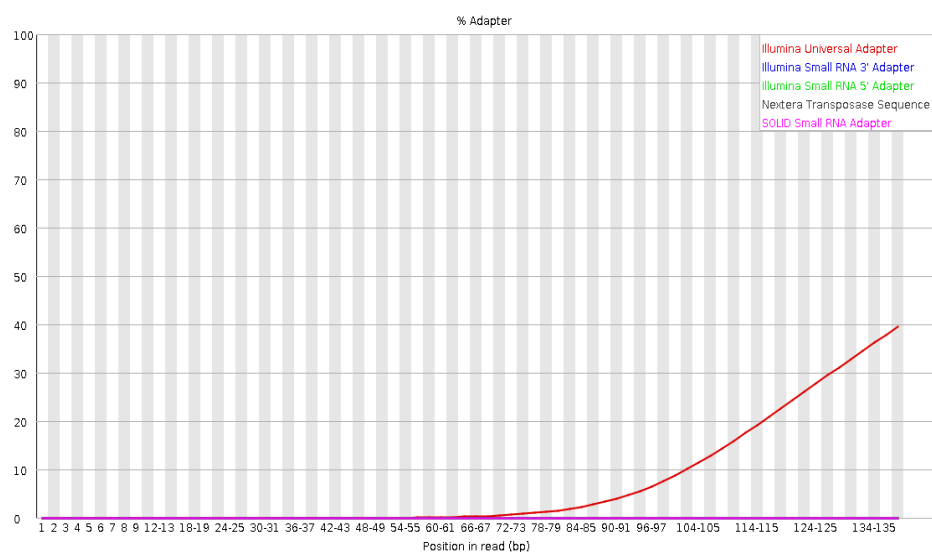


Figure B.98 This figure indicates the presence of adapter sequence contamination in sample C1-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

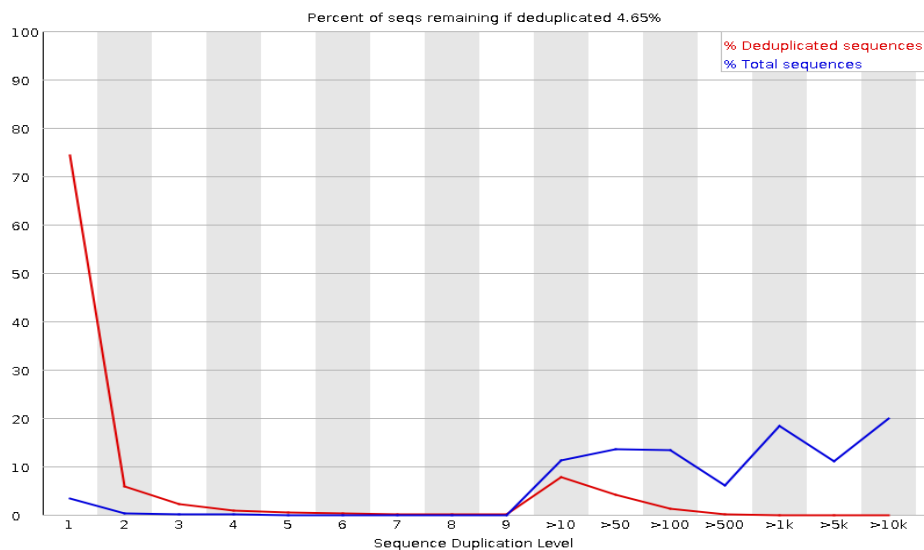


Figure B.99 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C1-R1. Here, 4.65% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

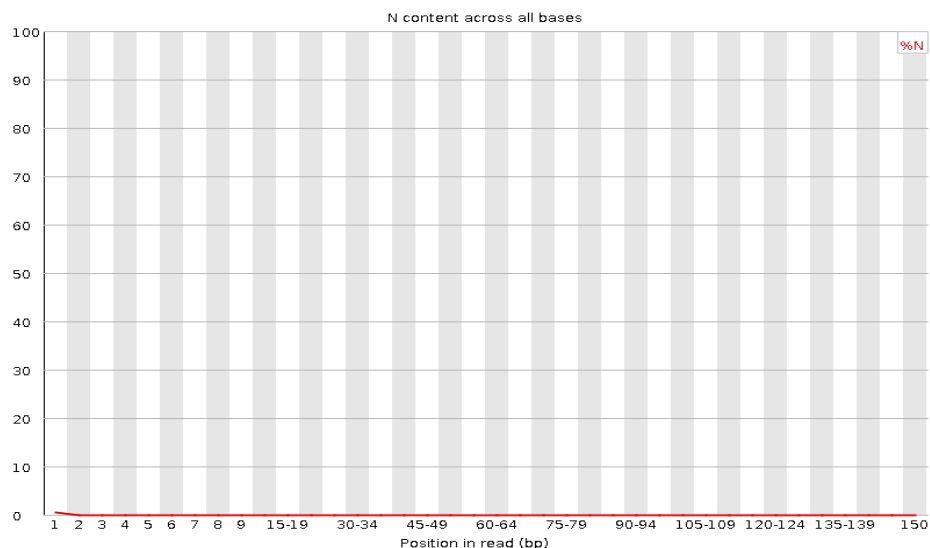


Figure B.100 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C1-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

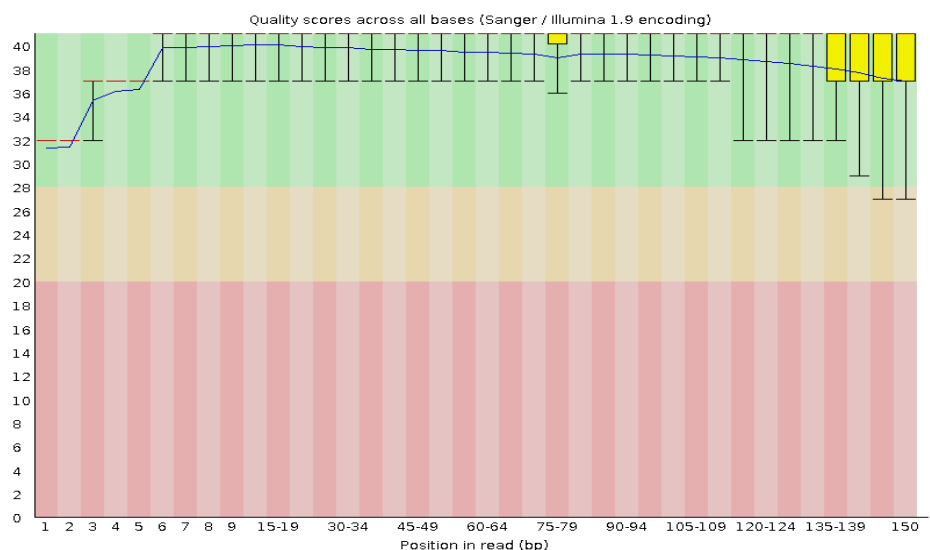


Figure B.101 Quality scores across the positions of the 150 bp reads for sample C1-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

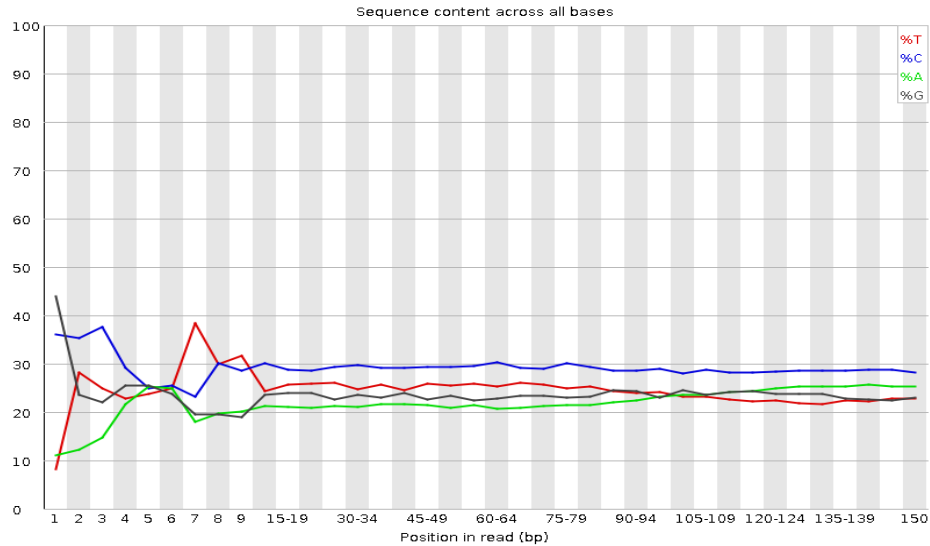


Figure B.102 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C1-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

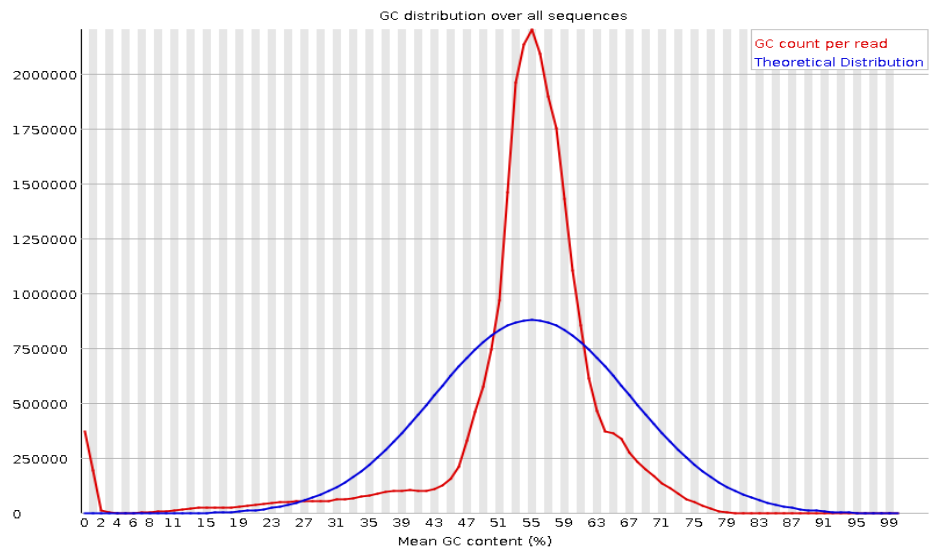


Figure B.103 The per sequence GC content for sample C1-R1. The x- and y-axes represent the %GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination.

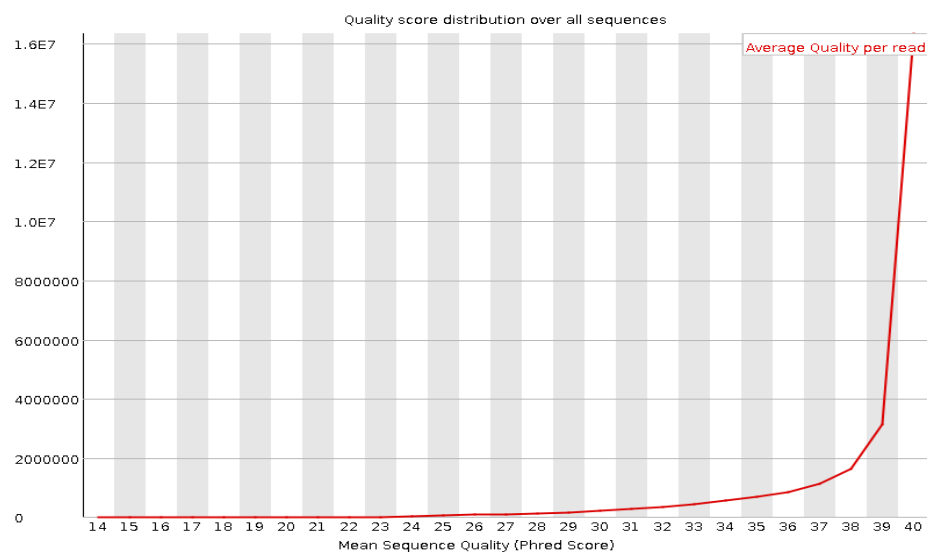


Figure B.104 The per sequence quality score for sample C1-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.



Figure B.105 The distribution of the sequence lengths for sample C1-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample C1-R2

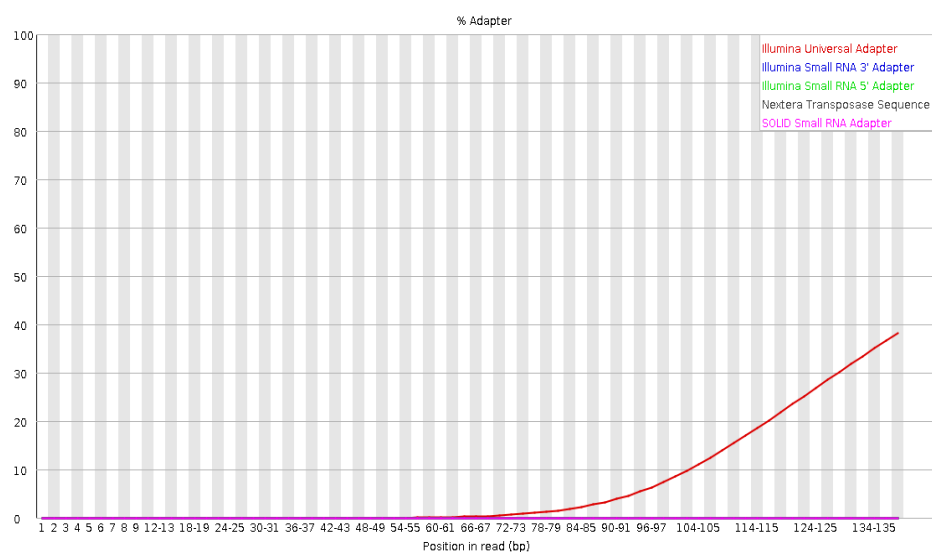


Figure B.106 This figure indicates the presence of adapter sequence contamination in sample C1-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

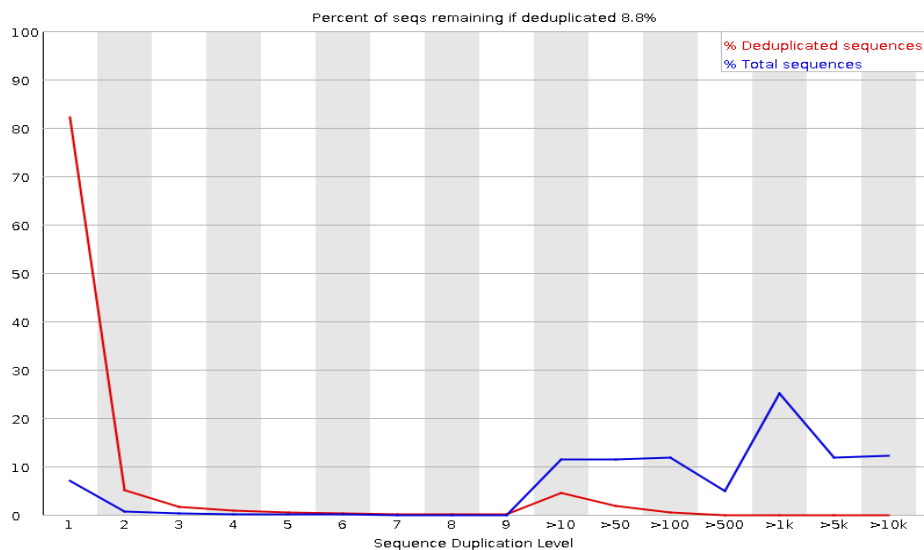


Figure B.107 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C1-R2. Here, 8.8% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

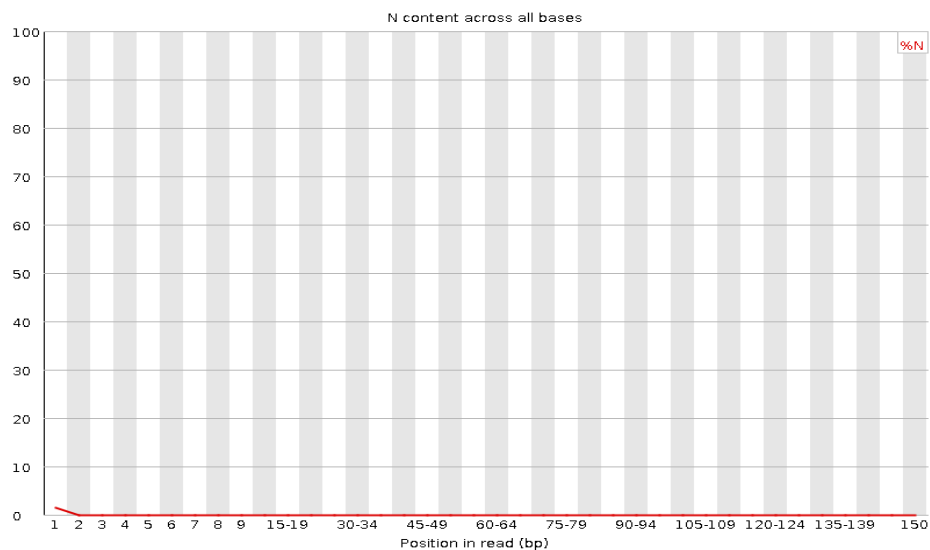


Figure B.108 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C1-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

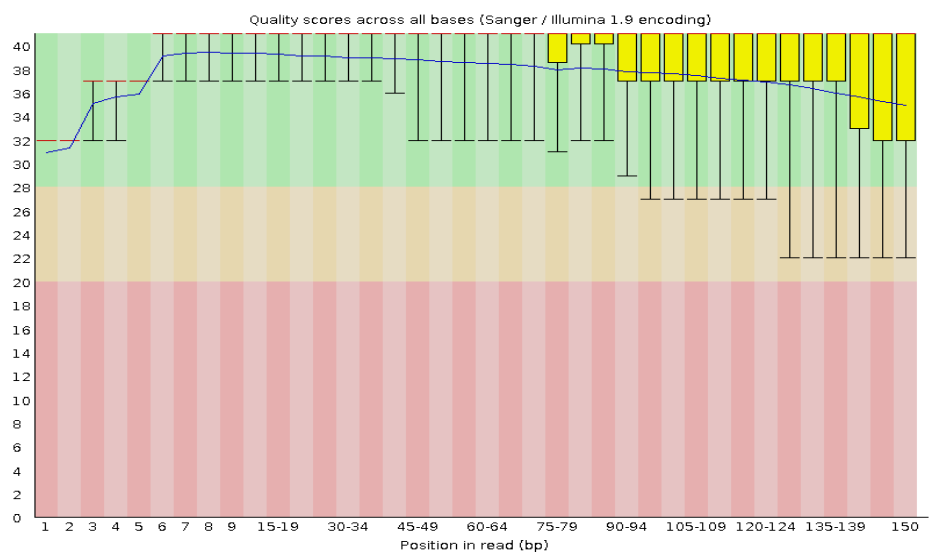


Figure B.109 Quality scores across the positions of the 150 bp reads for sample C1-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

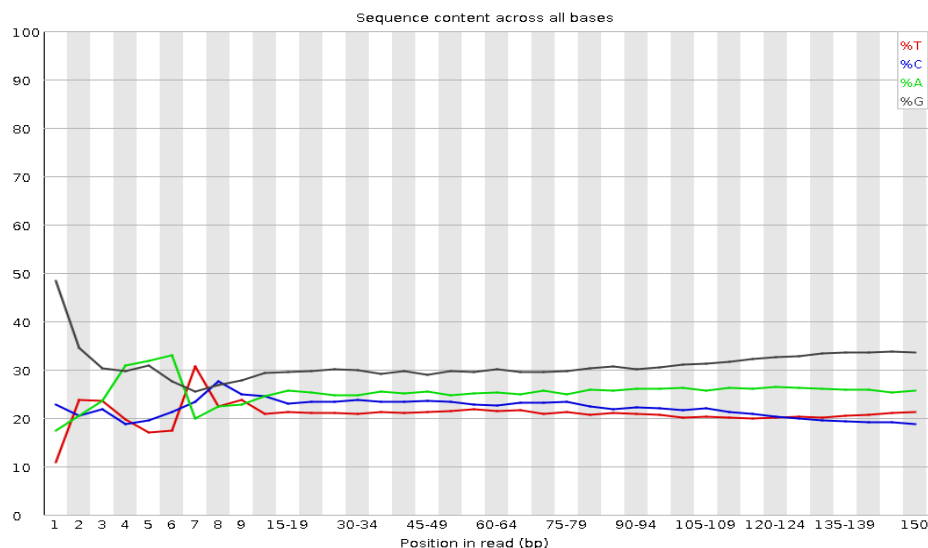


Figure B.110 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C1-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due to random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

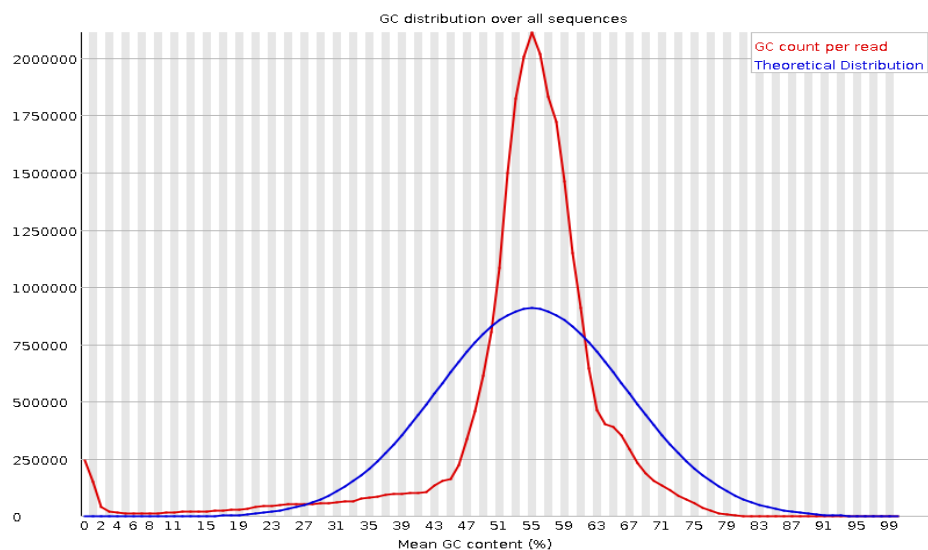


Figure B.111 The per sequence GC content for sample C1-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean % GC = 47 indicates bacterial contamination.

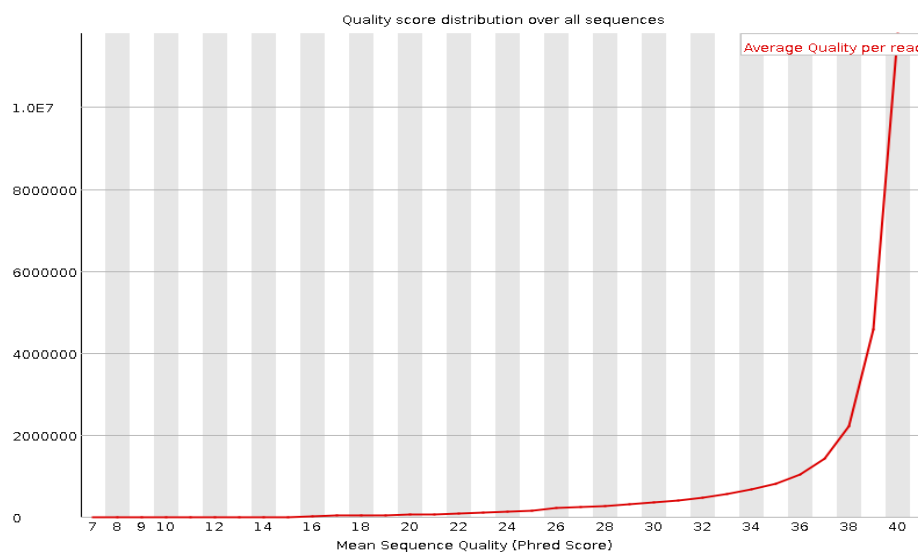


Figure B.112 The per sequence quality score for sample C1-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.

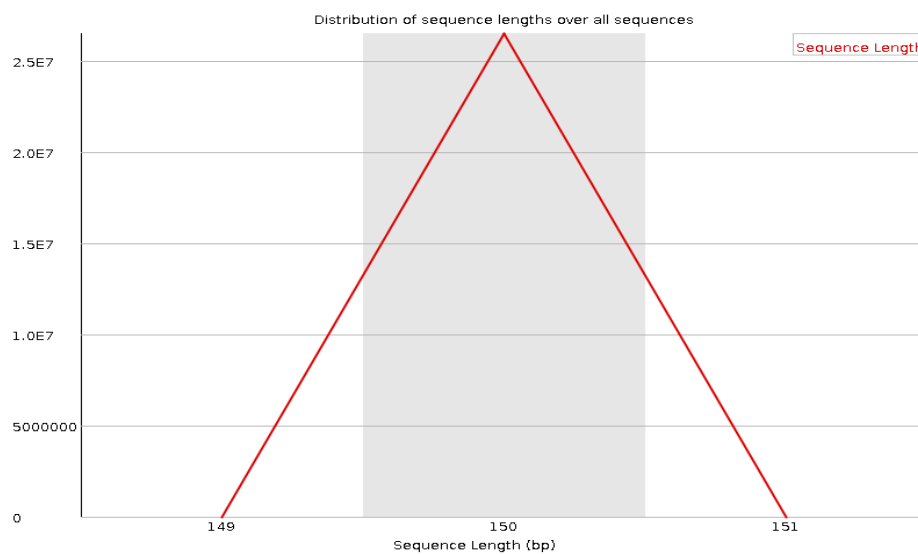


Figure B.113 The distribution of the sequence lengths for sample C1-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample C2-R2

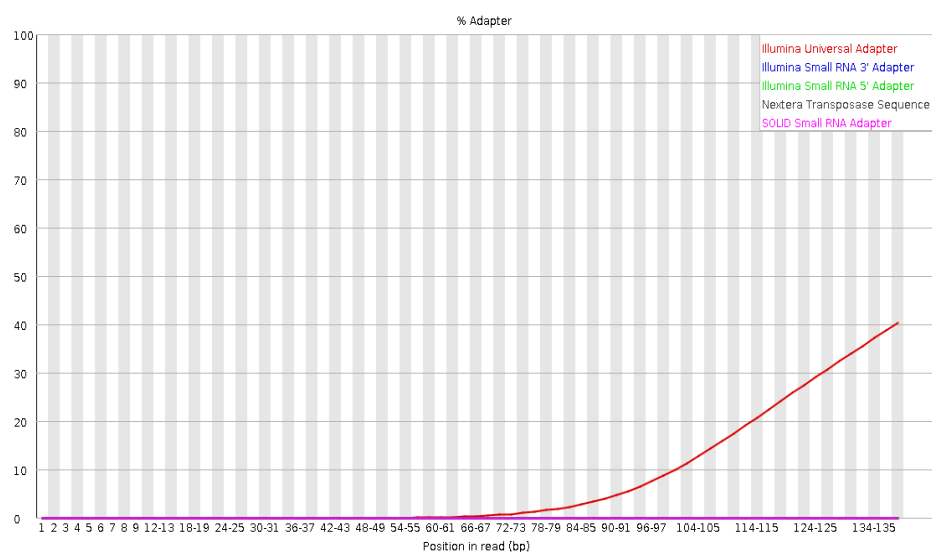


Figure B.114 This figure indicates the presence of adapter sequence contamination in sample C2-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

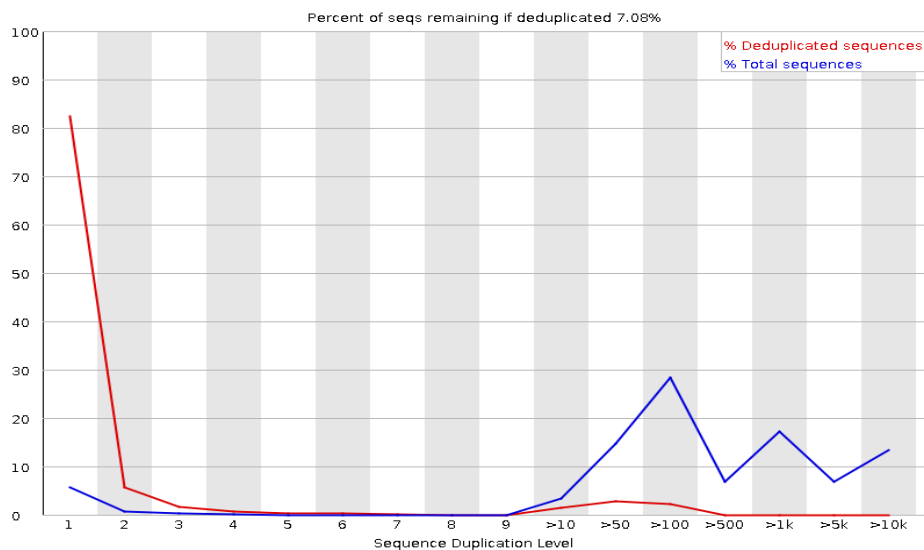


Figure B.115 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C2-R2. Here, 7.08% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

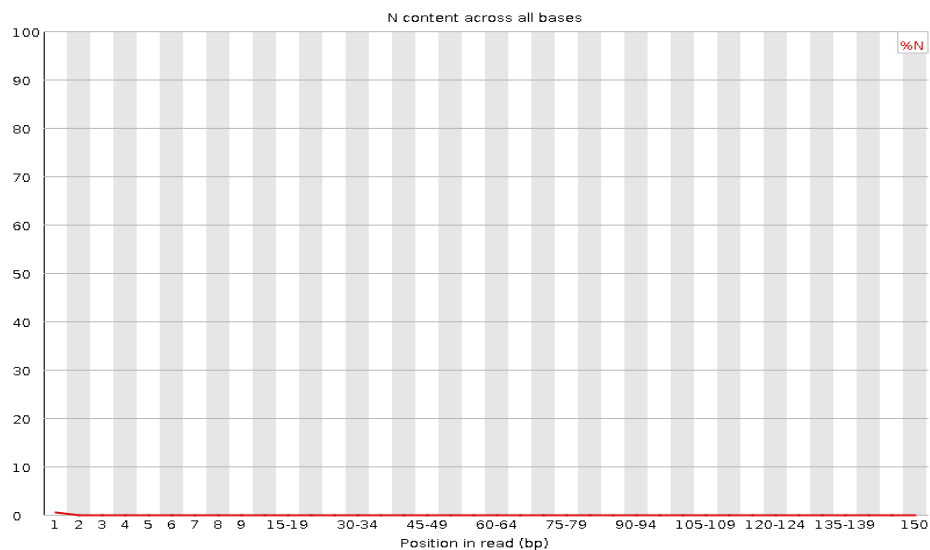


Figure B.116 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C2-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

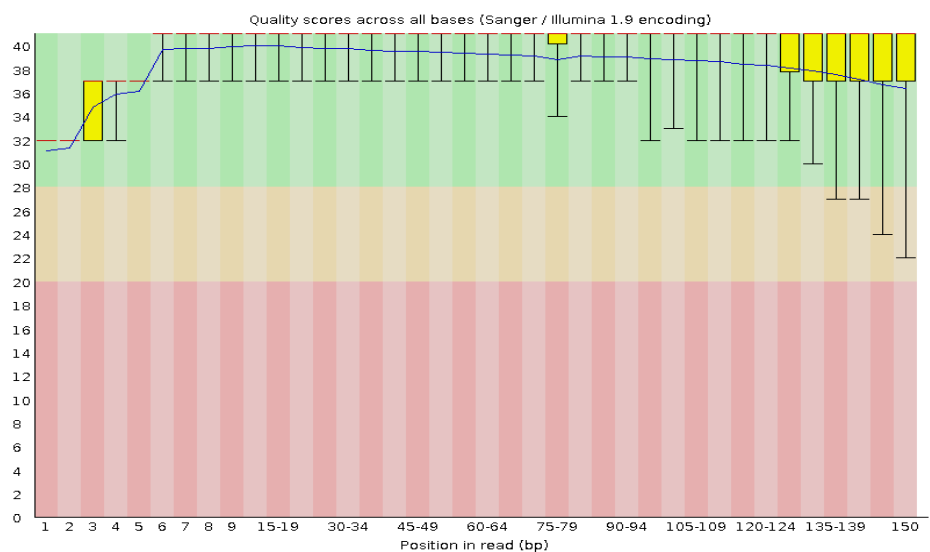


Figure B.117 Quality scores across the positions of the 150 bp reads for sample C2-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

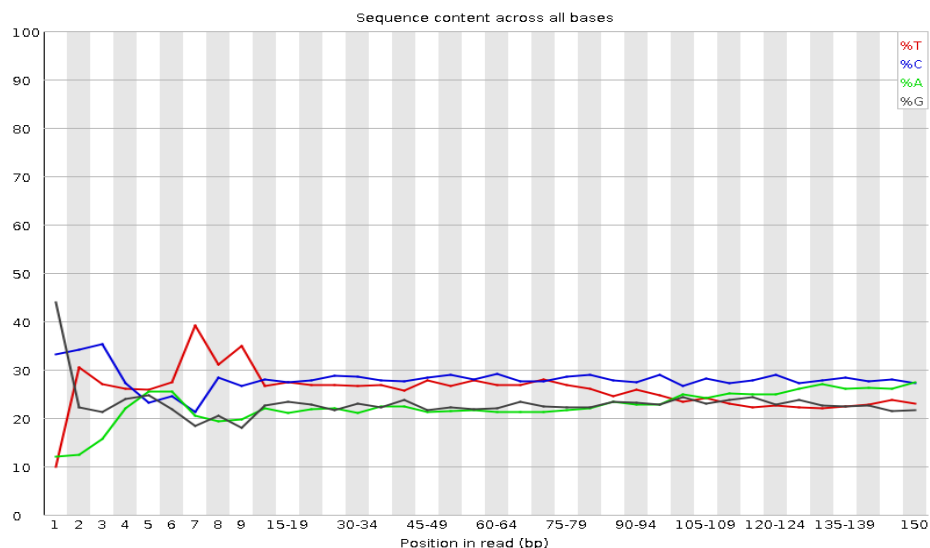


Figure B.118 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C2-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

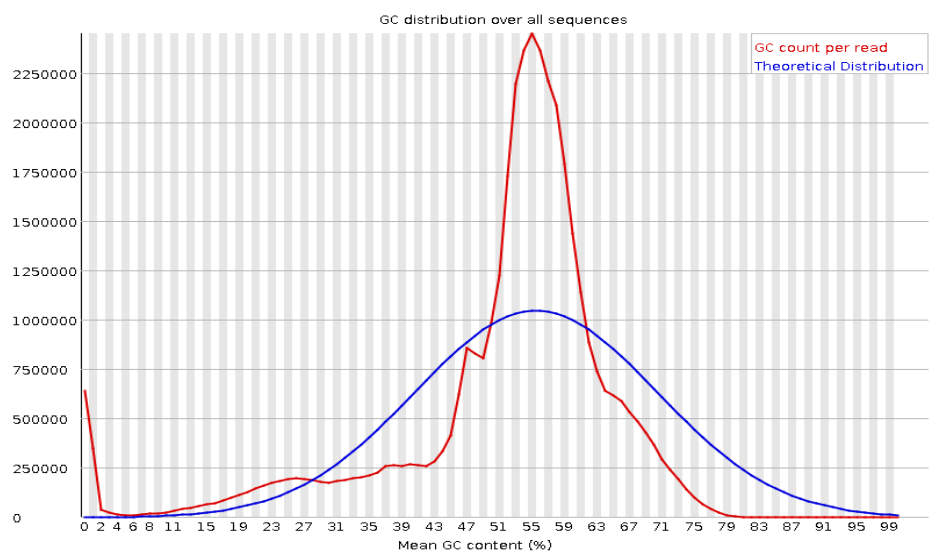


Figure B.119 The per sequence GC content for sample C2-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination. The sharp peak may indicate overrepresented bacterial reads.

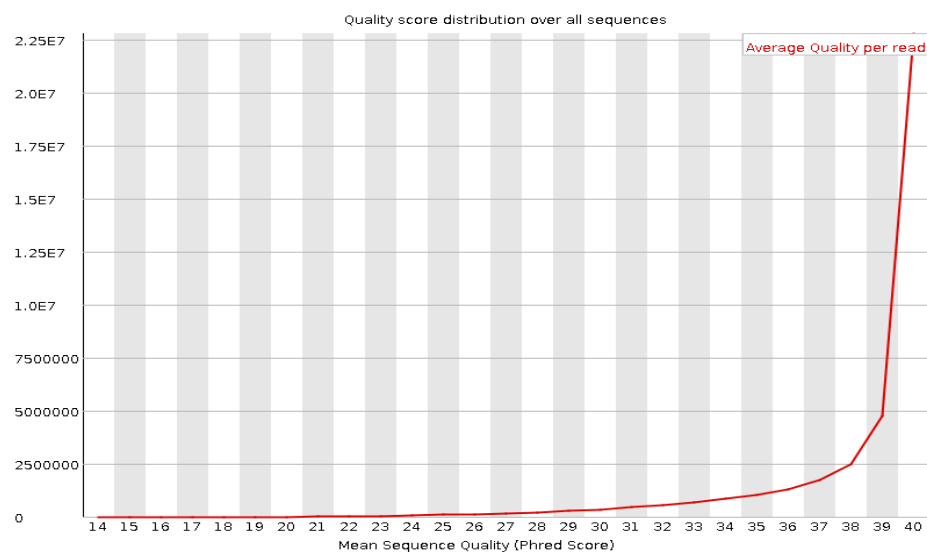


Figure B.120 The per sequence quality score for sample C2-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38.



Figure B.121 The distribution of the sequence lengths for sample C2-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample C3-R1

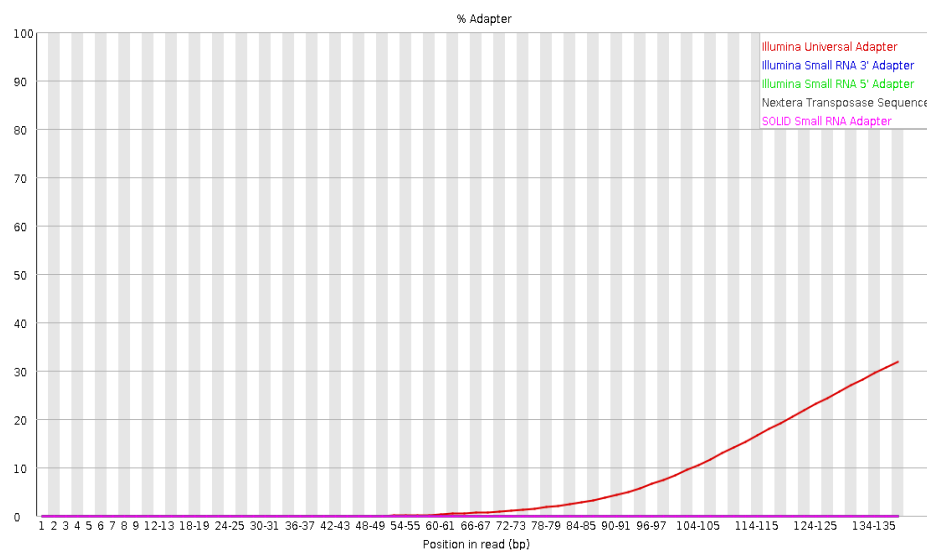


Figure B.122 This figure indicates the presence of adapter sequence contamination in sample C3-R1. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

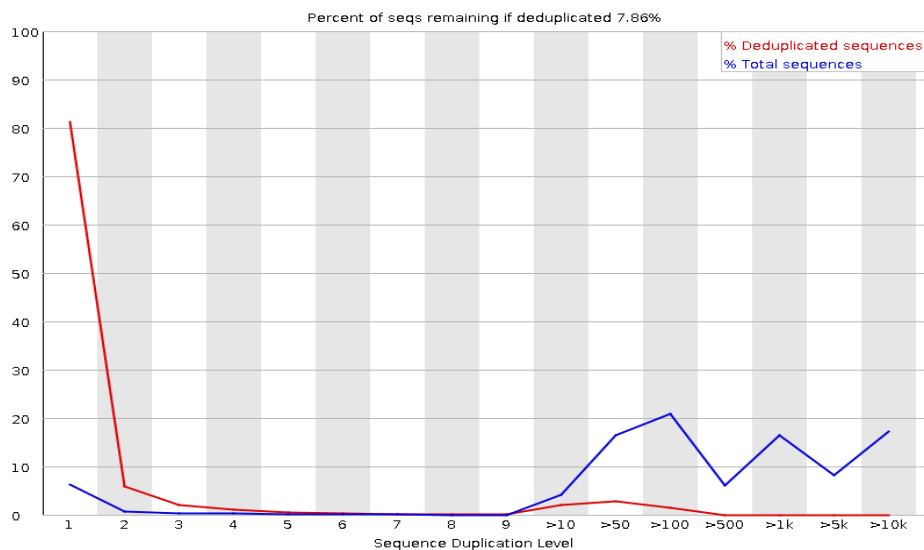


Figure B.123 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C3-R1. Here, 29.88% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

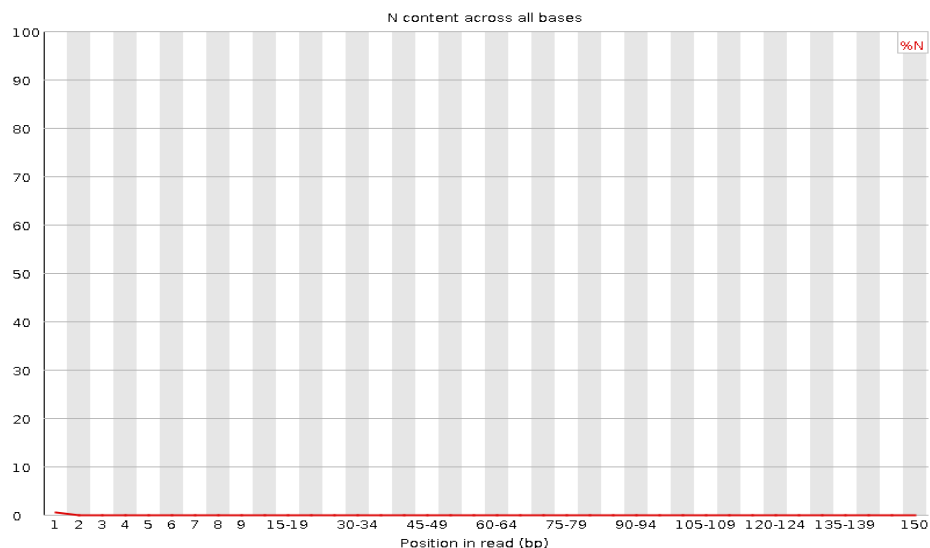


Figure B.124 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C3-R1. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

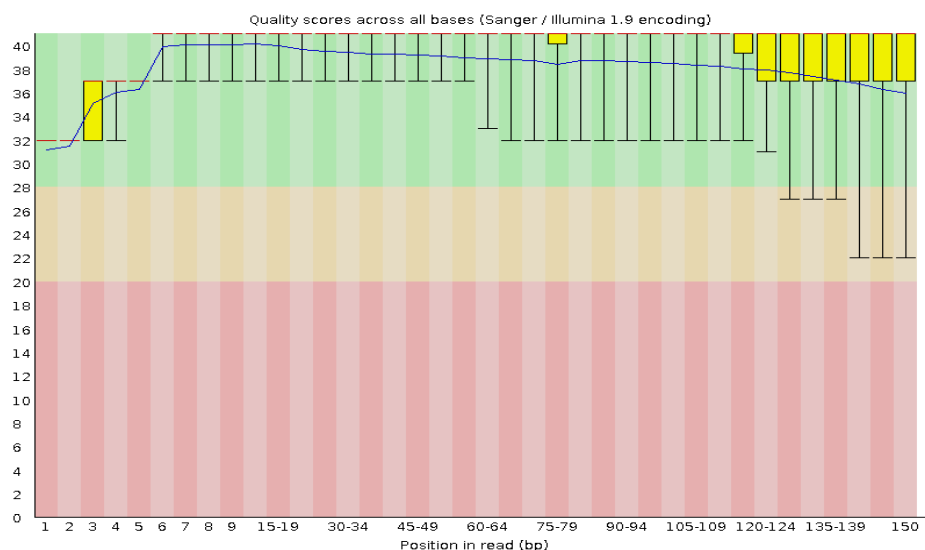


Figure B.125 Quality scores across the positions of the 150 bp reads for sample C3-R1. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

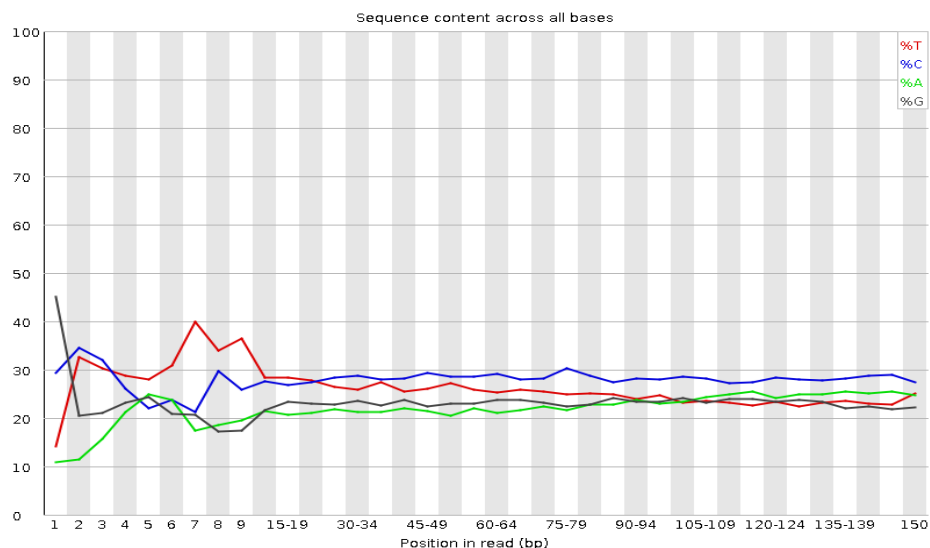


Figure B.126 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C3-R1. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

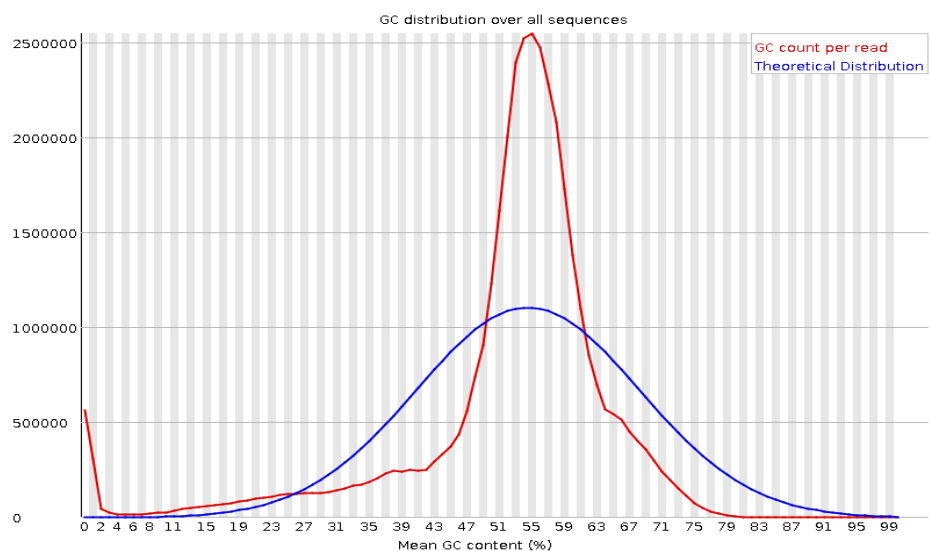


Figure B.127 The per sequence GC content for sample C3-R1. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination.

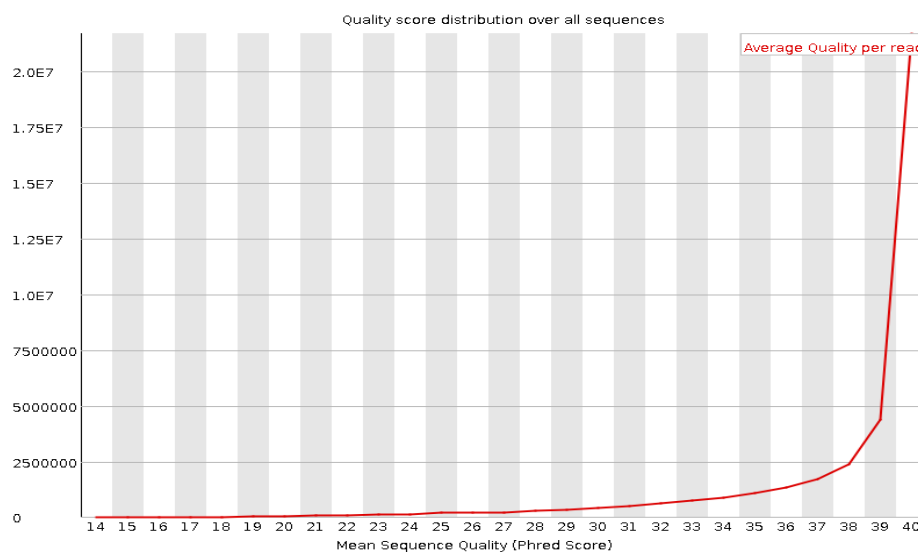


Figure B.128 The per sequence quality score for sample C3-R1. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.129 The distribution of the sequence lengths for sample C3-R1. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Sample C3-R2

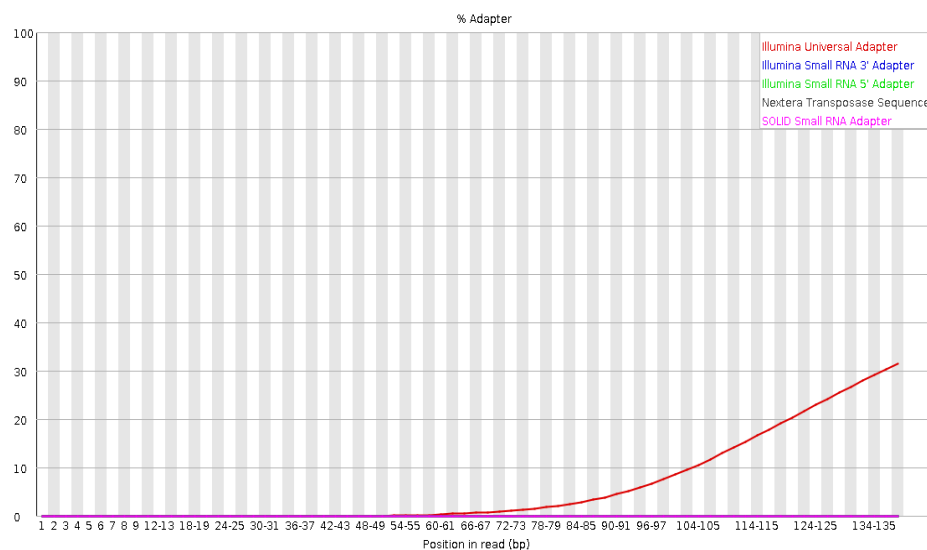


Figure B.130 This figure indicates the presence of adapter sequence contamination in sample C3-R2. The x- and y-axes represent the position within the 150 bp read and the percentage, respectively. The red line indicates the presence of the Illumina Universal Adapter starting at approximately bases 78-79 of 150, indicating the adapter trimming strategies should occur in the middle to 3' end of the reads.

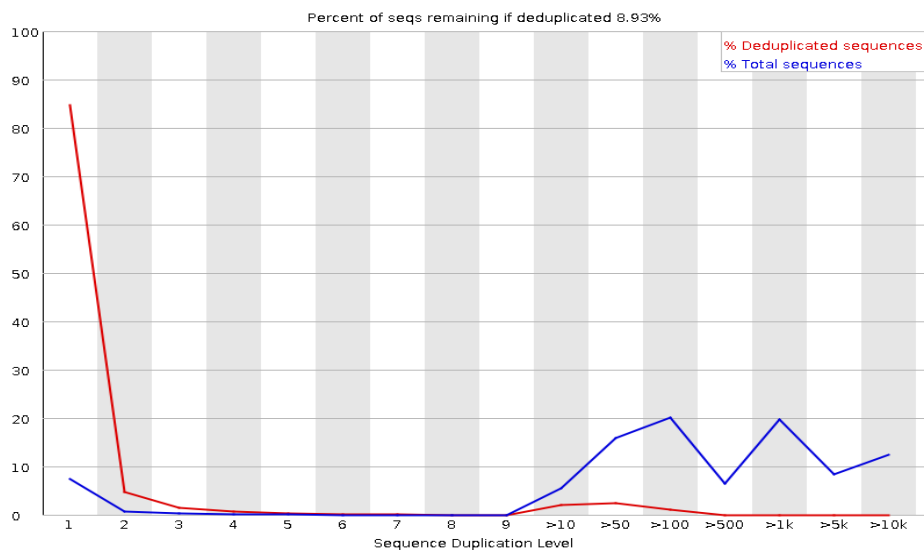


Figure B.131 The blue line represents the percent of the total sequences and the red line indicates the deduplicated (unique) sequences for sample C3-R2. Here, 8.93% of the reads were deduplicated. The peaks in the blue line, or the total sequences indicate the presence of contaminants or highly expressed transcripts under the conditions that were tested.

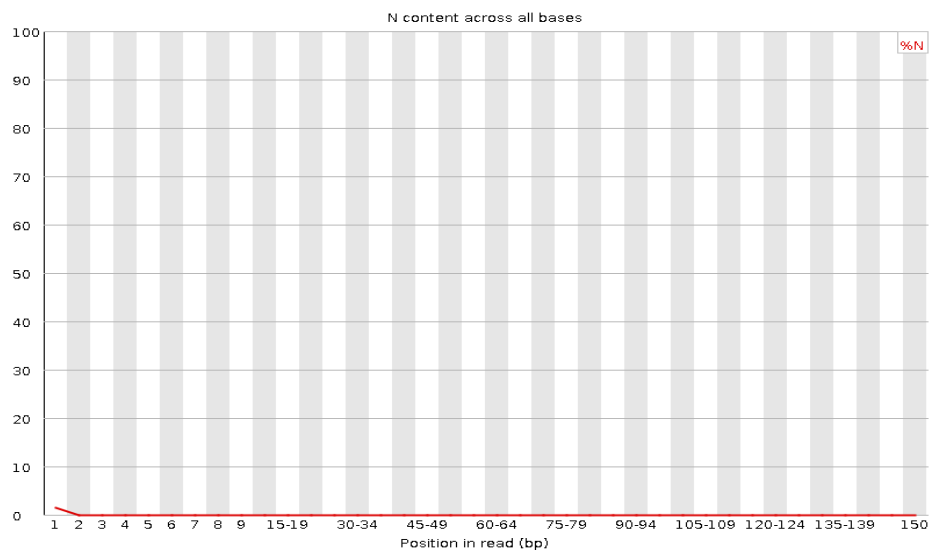


Figure B.132 The percentage of Ns and their positions across all bases in the 150 bp reads for sample C3-R2. The y-axis represents the percentage and the x-axis represents the position of the base in the read. The data presented here indicate that there are 0% Ns in the reads and all nucleotides are composed of A, T, C and Gs.

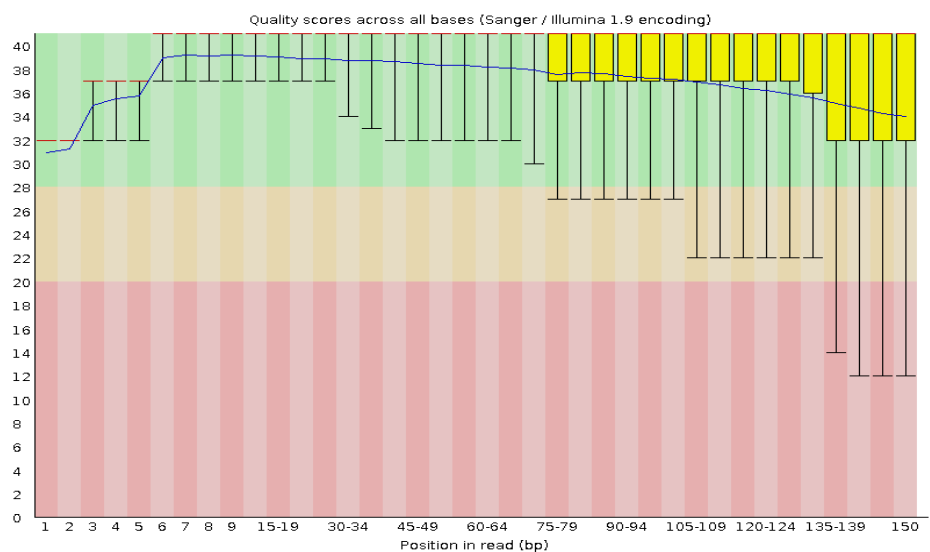


Figure B.133 Quality scores across the positions of the 150 bp reads for sample C3-R2. The x- and y-axes represent the quality scores and position within the read. The blue line represents the average quality of the bases at each position. The error bars represent 10 and 90% of the reads fall within that range. The yellow box represents 25-75% of the reads falling within that range.

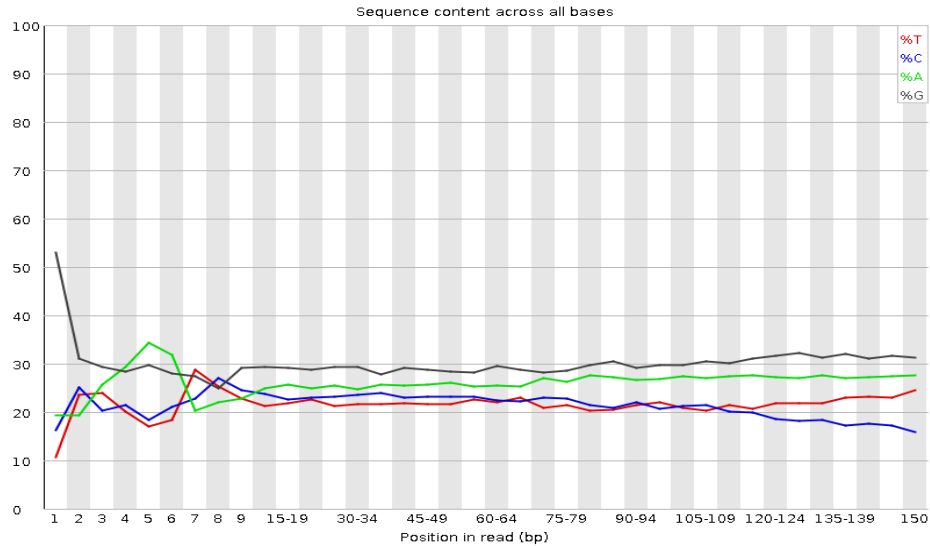


Figure B.134 The percentage of each nucleotide, A, C, T and G and their positions across the 150 bp reads for sample C3-R2. The x- and y-axes represent the position within the 150 bp reads and the percentage of each nucleotide. Ideally, each of the percent nucleotide lines should be parallel to each other. The erratic peaks in the beginning of the reads is due random hexamer ligation that imparts bias in RNA-seq libraries. While the bias does not affect the entirety of the reads, it enriches for k-mers at the 5' end of the reads.³

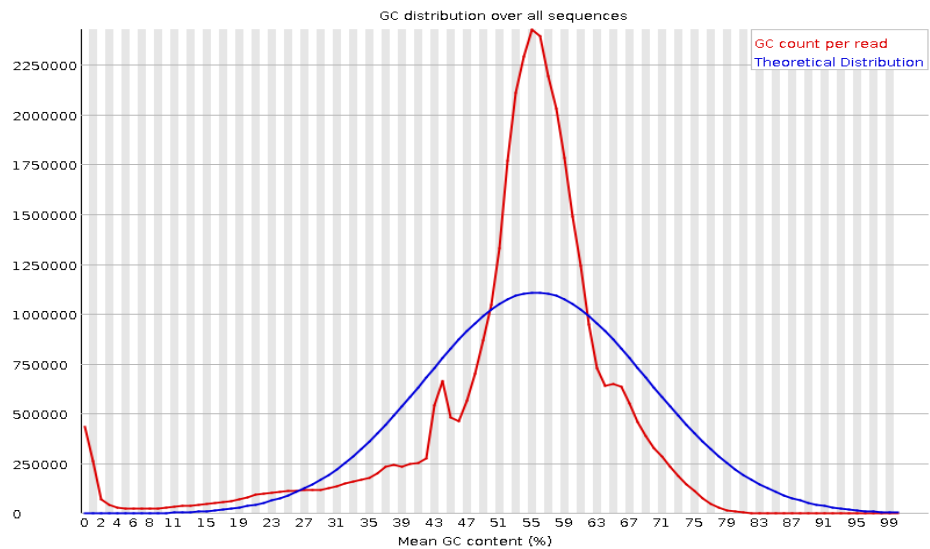


Figure B.135 The per sequence GC content for sample C3-R2. The x- and y-axes represent the % GC content per read and read counts, respectively. The blue line represents the GC content theoretical distribution and the red line represents the actual GC content for the 150 bp reads. The shift away from the mean %GC = 47 indicates bacterial contamination. The sharp peak may indicate overrepresented bacterial reads.

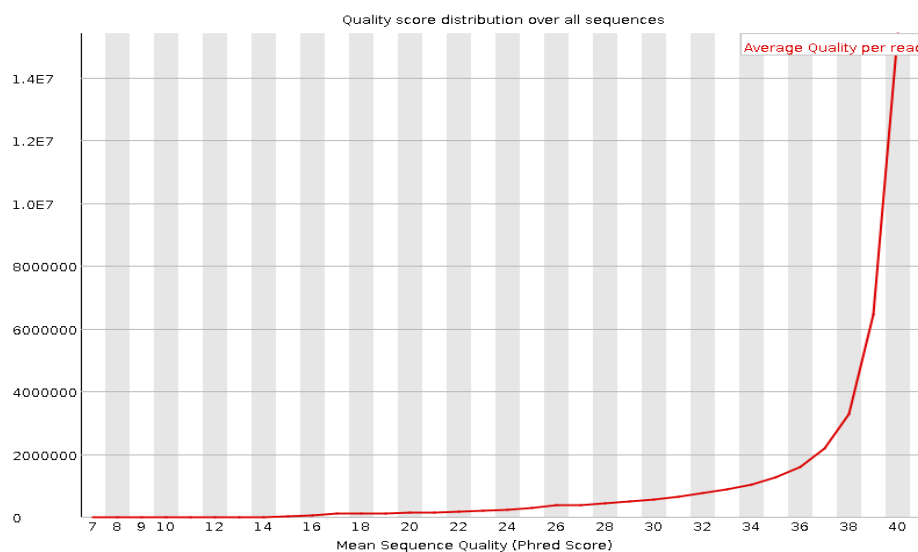


Figure B.136 The per sequence quality score for sample C3-R2. The x- and y- axes represent the mean quality (Phred Score) and the number of reads, respectively. The data here indicate that the majority of the quality scores were > 38 .



Figure B.137 The distribution of the sequence lengths for sample C3-R2. The x- and y-axes refer to the read sequence lengths and the number of reads with those lengths. The data here indicates that all reads were 150 bp.

Metabolic Pathways

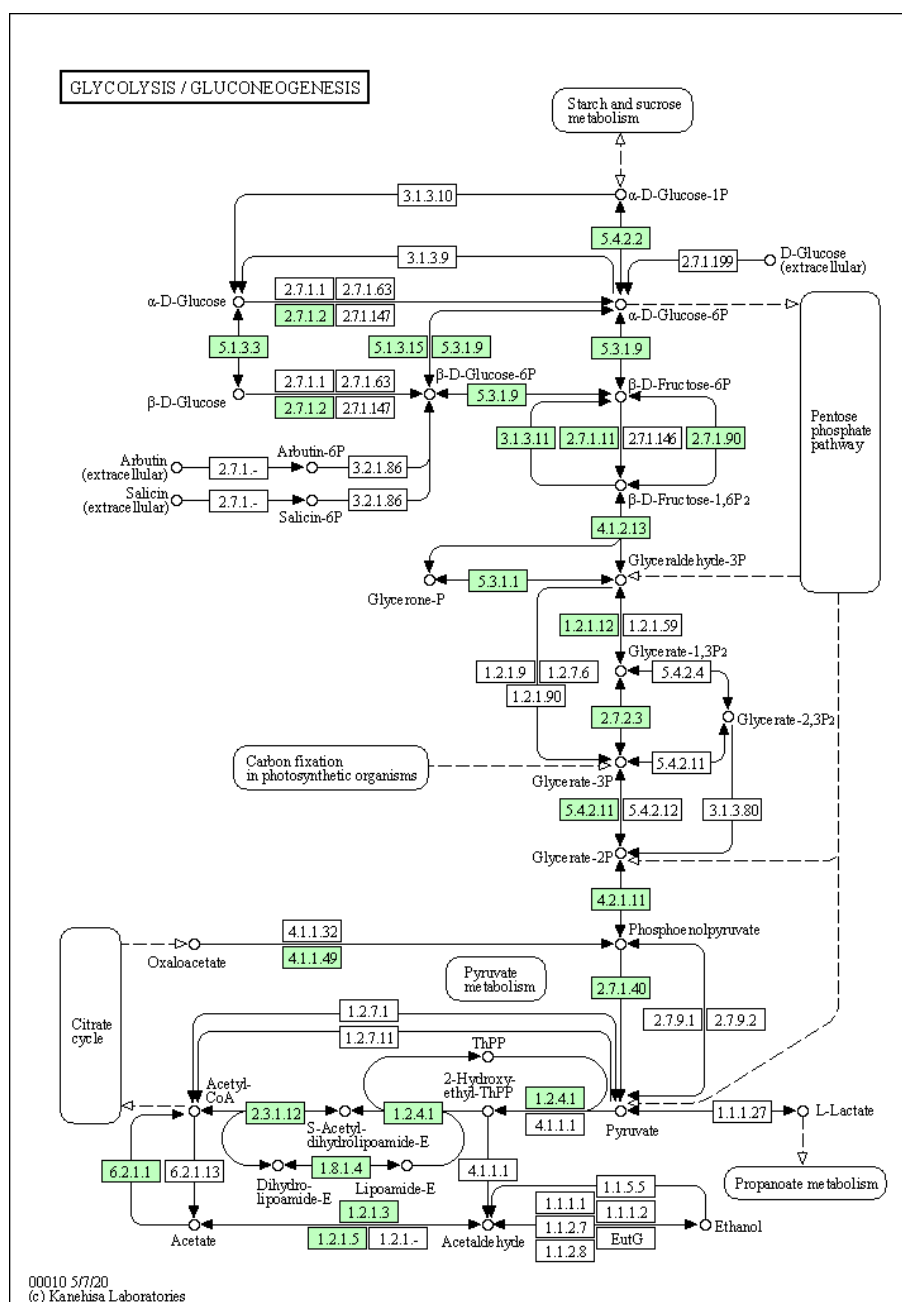


Figure B.138 The glycolysis/gluconeogenesis metabolic pathway with genes present (green) in the RGD-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

Figure B.139 The pyruvate metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

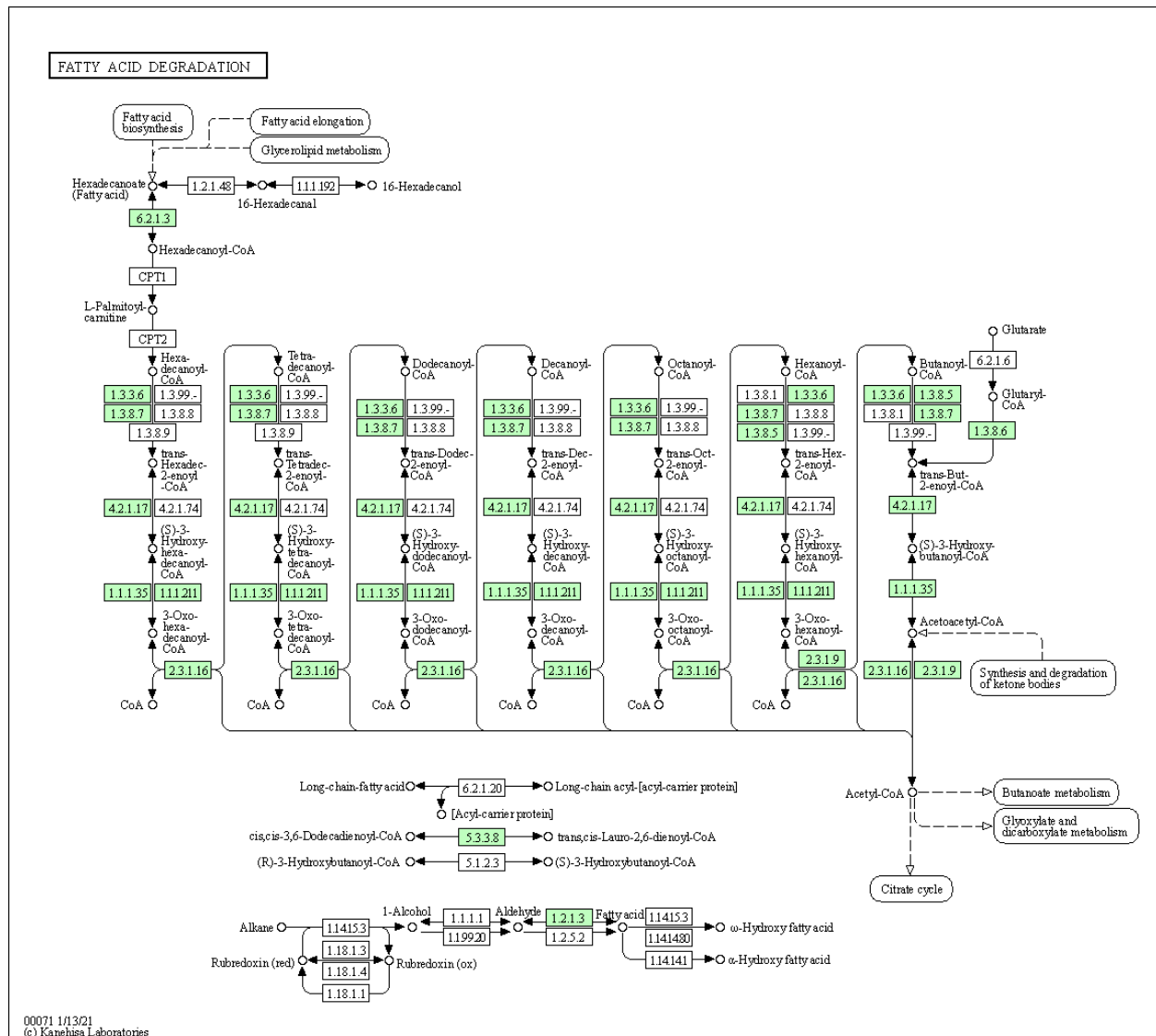


Figure B.140 The fatty acid degradation metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

Figure B.141 The glycerolipid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

Figure B.142 The α -linoleic acid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

Figure B.143 The arachidonic acid metabolic pathway with genes present (green) in the RGd-1 genome. The pathway was searched and populated using the online platform, KeggMapper.⁴

APPENDIX C

DETERMINING THE EFFECTS OF BLUE LIGHT ON THE RGD-1 GROWTH RATE

Introduction

Diatoms are responsible for approximately 40% of marine primary productivity and are known to assimilate 20% of global CO₂.²⁴⁷ As a necessary component for growth, diatoms fix atmospheric CO₂, which on a large-scale may help offset the increase in atmospheric CO₂ concentrations. In recent years, there has been renewed interest in using high lipid-producing algae strains for biodiesel production. As a near carbon-neutral technology, using algae for biofuel production may reduce further CO₂ emissions and help meet transport fuel demands. Outdoor raceway ponds are currently the most cost-effective method to grow algae for large scale biofuel production.^{5, 56, 248} However, to lower production costs, it is important to identify factors that contribute to enhanced growth and lipid accumulation.

Moll et al. (2014) found that diatom RGd-1 accumulated very high lipid concentrations; 30-40% (w/w) triacylglycerol (TAG) and 70-80% (w/w) biofuel potential (BP).⁴⁹ Due to its ability to produce such high concentrations of lipids, it was selected for whole-genome sequencing. Currently, there are only two other diatoms with sequenced, published and publicly available genomes.

RGd-1 was observed to differentiate into multiple cell sizes and cell morphologies (Figure C.1). One hypothesis that may account for different cell phenotypes is a lack of blue light in laboratory light systems. A recent study discovered novel, diatom specific genes in the blue light cryptochrome/photolyase family (CPF) in the *Phaeodactylum tricornutum* genome, which was found to act as transcriptional regulators for cyclin expression.^{249, 250} Overexpression of PtCPF1 was found to activate expression for blue light-induced genes that are known to be involved in the cell cycle and DNA repair^{250, 251}. This suggests that CPF1 may sense changes in environmental light conditions and may play a role in signaling the completion of the cell cycle

with sufficient blue light. Another newly described, stramenopile specific blue light-absorbing group of photoreceptors have also been identified as aureochromes^{252, 253}. In *P. tricornutum*, AUREO1a is a transcription factor regulating the diatom specific cell cycle protein, dsCYC2, which controls the G1-S phase of the cell cycle^{252, 253}. AUREO1a is a transcription factor that regulates the diatom specific cell cycle protein, dsCYC2, which controls the G1-S phase of the

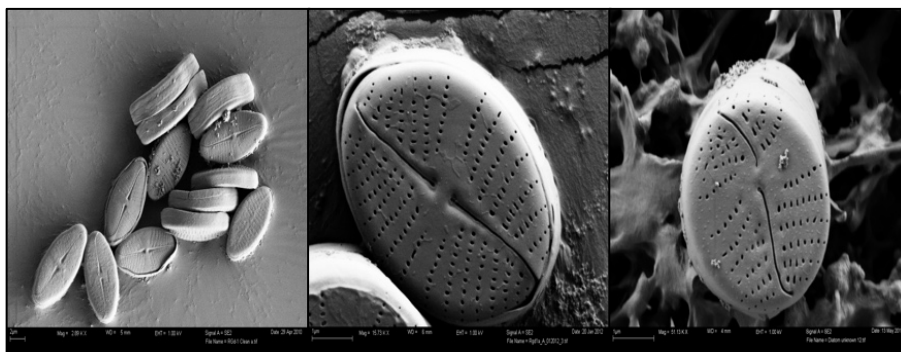


Figure C.1 RGD-1 cellular morphologies as imaged using field emission – scanning electron microscopy. The RGD-1 morphology in 2009 (left), and different cell morphologies in 2014 (middle and right).

cell cycle^{252, 253} and CPF1 may sense changes in environmental light conditions and may play a role in signaling the cell cycle as a result. Both AUREO1a and CPF1 have been identified in the RGD-1 genome. At least 18 significant hits with e-values $1e-5$ or lower, and %ID of at least 80% were found when aureochromes from *P. tricornutum* were BLAST searched against the current RGD-1 genome assembly.

At present, the RGD-1 growth rate is slower than desired (~29h doubling time). To improve growth rates, other strategies have been employed such as changing the Si, Fe and As concentrations as well as the light intensity. Here we have focused on varying the intensity of blue light (350-500 nm) to observe its effects on the RGD-1 growth rate. As shown in the results, the light spectrum produced from fluorescent algae grow lights is significantly deficient in blue light as compared to the natural light in Yellowstone National Park, where RGD-1 was isolated.

It was hypothesized that varying the blue light intensity would affect progression through the cell cycle which may alter the growth rate. However, to the best of our knowledge, no studies have elucidated this effect on diatom growth. While the studies described above investigated the effects of different light wavelengths on circadian rhythm specific gene expression, there is a current gap in the knowledge of the effects of varying blue light intensities on diatom growth rates.

Background

Cryptochromes and photolyases are flavoproteins with blue light (350-500 nm) sensitive receptors that regulate gene expression and catalyze the repair of UV damaged DNA (repair of cyclobutane pyrimidine dimers or 6-4 photoproducts), respectively.^{254, 255} A novel cryptochrome/photolyase family protein (CPF) was recently identified in the *P. tricornutum* genome.²⁵¹ Blue light is prevalent in the upper layers of the water column where diatoms most commonly reside, and it is one of the few wavelengths able to penetrate to greater depths in the water column.^{251, 256} *Arabidopsis* has been found to have seven blue light receptors, of which, cryptochromes (CRY1 and CRY2) have been found to have activity in hypocotyl elongation, floral initiation, the circadian rhythm, chloroplast development, guard cell development, stomata opening.²⁵⁷ Other possible functions including root development, programmed cell death, magnetoreception, seed dormancy, pathogen responses, and the cell cycle.²⁵⁸ For the pennate diatom, *Navicula*, blue light-induced dense, evenly-distributed lawns of cells and increased the growth rate more than three times compared to growth on other wavelengths.²⁵⁶ For *P. tricornutum*, *PtCPF1* overexpression resulted in altered expression for genes associated with photoinhibition, thus decreasing photosynthetic efficiency which would limit growth and lipid accumulation.^{251, 253}

In animals, the CPFs act as transcriptional activators controlling the circadian rhythm.²⁵¹ The CPFs identified in both *T. pseudonana* and *P. tricornutum* are more closely related to animals than plants. This is expected given the evidence for secondary endosymbiosis making diatoms phylogenetically more related to animals than plants.^{36, 204, 255} Previous studies have investigated the role of animal cryptochromes (aCRY) in the green alga, *Chlamydomonas reinhardtii* and genome-wide changes in gene expression for *T. pseudonana* with regards to light at different times of the day.^{204, 255} Su et al. 2015 investigated the effects of different wavelengths of light on the growth rate and frustule size for the diatom, *Coscinodiscus granii*. Blue and red light wavelengths were found to produce the fastest growth rates and lower intensities of each wavelength resulted in morphological differences in the diatom frustule.^{259, 260} While this study looked at the effect of two different intensities using six different light-emitting diodes (LEDs) targeting the wavelengths of interest, they did not tightly control the wavelengths used in their studies or look at the transcriptomic effects of these conditions.

Several diatom specific cyclins, proteins associated with the progression of the cell cycle, have been identified.^{67, 250, 261} Specifically, *P. tricornutum* and *T. pseudonana* have 24 and 52, respectively eleven of which were determined to be diatom specific.²⁶¹ Bowler et al. 2008 suggested that this large number of cyclins may be attributable to the unique diatom life cycles, affecting diatom cell size.³⁵

A study by Kafri et al. 2013 found that larger, faster growing tissue culture cells, halted progression throughout the cell cycle at the G1/S checkpoint until the smaller, slower growing cells caught up.²⁶² This allowed all cells in a culture to progress through the remainder of the cell cycle uniformly at the same rate. A deficiency in blue light may alter progression through the cell cycle, resulting in a decreased growth rate.

The hypothesis for this work was that there is an inverse relationship between blue light intensity and the RGd-1 growth rate. The approach was to change the intensity of the blue light spectrum only (zero, low, medium and high intensity) while maintaining similar intensities for all other wavelengths and overall photosynthetically active radiation (PAR).

From this study, the effect of blue light on RGd-1 growth, cellular morphology, and lipid productivity was determined. It was expected that we would see an increased growth rate with increased blue light intensity and potentially improve our ability to grow this high lipid-producing diatom. Stabilizing phenotypic traits such as growth rate and lipid accumulation is vital for improving the viability of algal biofuels.

Methods

All experiments were performed in temperature-controlled tubular photobioreactors using 1.25 L of sterile alkaline growth medium (modified Bold's Basal Medium supplemented with B₁₂, S3 vitamin solution and titrated to pH 8.7), grown with a 14:10 light/dark (L:D) cycle at 27°±1°C.⁴⁹ Growth studies were performed using LEDs in the MSU phototrophic growth lab. Each experiment was maintained at 22% light intensity using a custom bank of light-emitting diodes (LED) (25.4 cm wide, 120.65 cm long). Three light filters were used blue (filter no. 384), light yellow (filter no. 313), yellow (filter no. 6) and a control (no filter) (Table C.1) (Rosco) were wrapped around the outside of the photobioreactor tubes to adjust the blue light intensities. All conditions remained consistent except for the filter type. Each culture was grown in duplicate with filters wrapped around each individual 1.25 L photobioreactor for a total of eight tubes. The blue light intensities were varied using filters (Rosco) that filter out blue light while leaving the other wavelengths at a consistent intensity.

Table C.1 Growth conditions (filter types) and the measured PAR passing through the filter measured with a spectroradiometer (Ocean Optics).

Filter	PAR ($\mu\text{mole photons m}^{-2} \text{ s}^{-1}$)
Control	1.97e^{-4}
Light Yellow (#313)	1.45e^{-4}
Yellow (#6)	1.70e^{-4}
Blue (#384)	1.72e^{-5}

To evaluate the effect of blue light on growth, the following measurements were performed daily as outlined by Moll et al. 2014⁴⁹: direct cell counts, pH, chlorophyll and Nile Red fluorescence intensity (rfu). To quantify the growth, a minimum of 100 cells were counted from each sample using a hemacytometer (Reichert). The sample pH was measured using a standard benchtop pH meter (Accumet). The lipid accumulation was measured daily by staining the cultures with Nile Red (0.25 mg/mL suspended in 20% DMSO) and using a microplate reader (BioTek Synergy). Final dry cell weights (DCW) were assessed at the end of the experiment by filtering 25 mL of each culture with pre-weighed F/F Glass Microfiber Filters (Whatman). Samples were re-weighed after drying at 60°C after approximately 24 and 48 hours. Once experiments reached maximum Nile Red fluorescence intensity, cultures were harvested, and dry cell weight was measured. Chlorophyll was extracted using acetone and measured on a plate reader (Bio-Tek) at 632, 652 and 665 nm to quantify chlorophylls a and c.^{263, 264}

It was expected that when the blue light intensity was changed, the PAR would also change. The filters were chosen to avoid this problem as much as possible. The PAR and light spectrum produced by each filter was measured using a spectroradiometer (Ocean Optics).

Results and discussion

The overall blue light emitted from the MSU fluorescent grow lights was significantly lower compared to the blue light produced by the natural sunlight at Witch Creek, where RGd-1 was isolated (Figure C.2). By changing the blue light intensity, it was hypothesized that there would be a significant difference in growth rates between the different conditions.

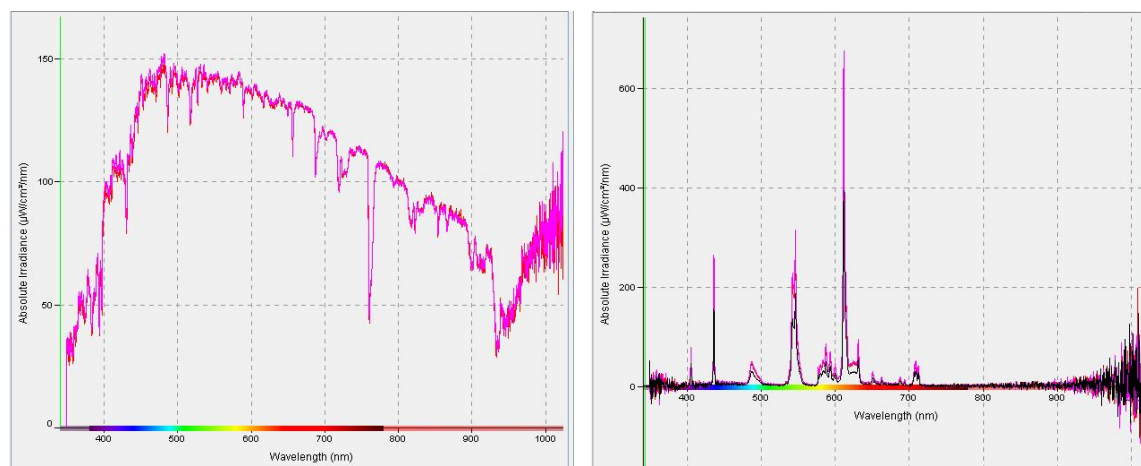


Figure C.2 The light intensity at Witch Creek (late morning August 2012, left) and laboratory fluorescent grow lights (right). Two measurements were taken in the field (1807 and 1828 $\text{uW cm}^{-2} \text{nm}^{-1}$) and three measurements were taken for the MSU lab light systems (421, 412 and 379 $\text{uW cm}^{-2} \text{nm}^{-1}$).

Four conditions were tested; a no filter control, and three different blue light filters with different blue light intensities (Table C.1). The light yellow filter had the lowest blue light intensity, the yellow filter had a higher blue light intensity and the blue filter had the highest blue light intensity but was deficient in the other wavelengths greater than approximately 550 nm. The PAR was substantially lower for the blue light filter (#384) compared to the other two filters and control (Table C.1).

The light emission spectra for the four conditions that were tested are shown in Figures 3-6; the no filter control (Figure C.3), light yellow (#313) (Figure C.4), yellow (#6) (Figure C.5) and blue (#384) (Figure C.6). It is important to note that the y-axis is different for the conditions depending on the highest intensity. The no-filter control had the highest intensity and coverage for wavelengths greater than 500 nm (Figure C.3). Compared to the light yellow and yellow filters (Figures C.4 and C.5), the no-filter control also had a higher blue light intensity. The blue light filter had the highest blue light intensity but was deficient in wavelengths exceeding approximately 550 nm (Figure C.6).

Table C.2 shows that each of the four conditions reached similar final cell concentrations at approximately 1.0×10^6 cells mL⁻¹. However, the control was slightly higher at 1.45×10^6 cells mL⁻¹ $\pm 1.41 \times 10^4$ cells mL⁻¹, which could be due to a reduced PAR inside the tubes with filters (Table C.1, Figure C.7). Further, the doubling times reflected this as well. The light yellow filter, with the lowest blue light intensity (Figure C.4), resulted in the fastest doubling time at 27.04 ± 12.12 h (Table C.2). However, it was not statistically different from the control condition that grew at a doubling time of 38.49 ± 10.55 h. (Figure C.7). When RGd-1 was grown with the yellow filter that had a greater intensity of blue light that passed through the filter (Figure C.6), the doubling time decreased compared to the light yellow and control conditions at 40.23 h (Table C.2, Figure C.7). One of the photobioreactor tubes started clumping so only one of the two tubes was used for analysis.

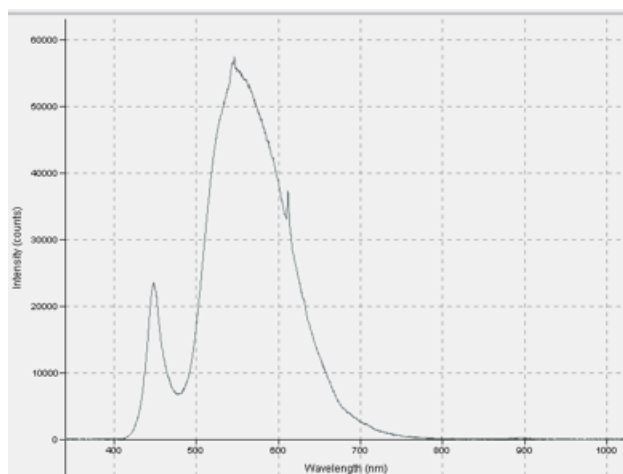


Figure C.3 The control (without a color filter) spectroradiometer measurement at 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.

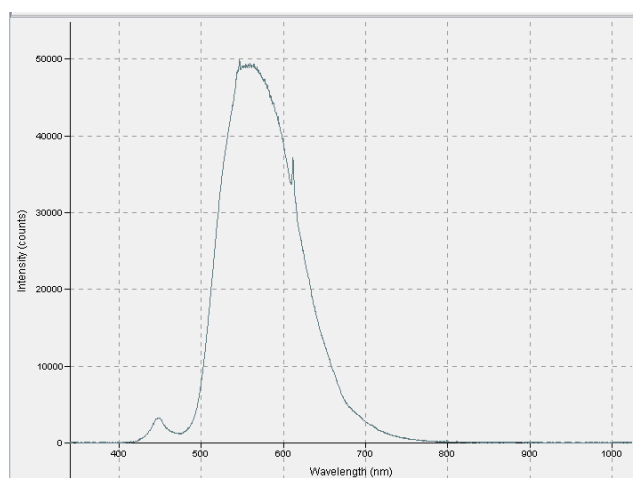


Figure C.4 Spectroradiometer measurement (Ocean Optics) for the Rosco filter #313 (light yellow) with low blue intensity using 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.

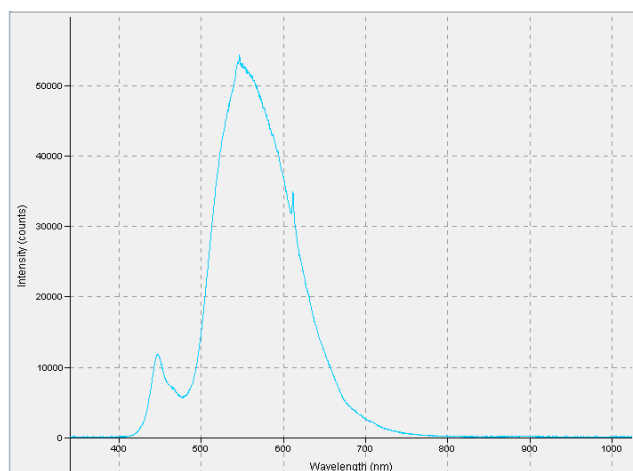


Figure C.5 Spectroradiometer measurements (Ocean Optics) for the Rosco filter #6 (yellow) high blue intensity using 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.

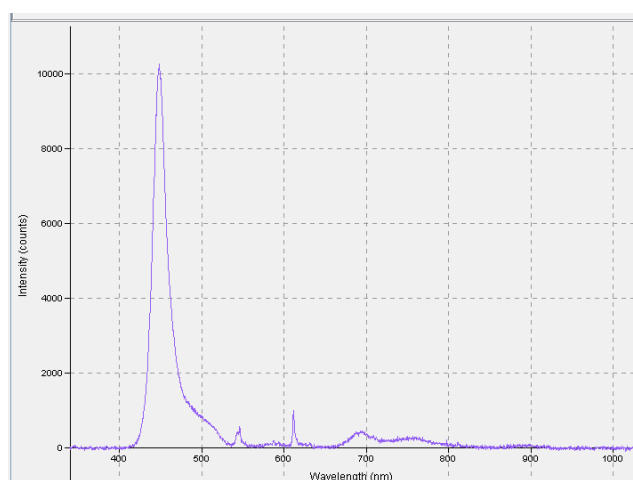


Figure C.6 Spectroradiometer measurements (Ocean Optics) for the Rosco filter #384 (blue) very high blue intensity, and low intensity for other wavelengths and 22% LED intensity. The spectroradiometer measurements were taken inside empty photobioreactor tubes.

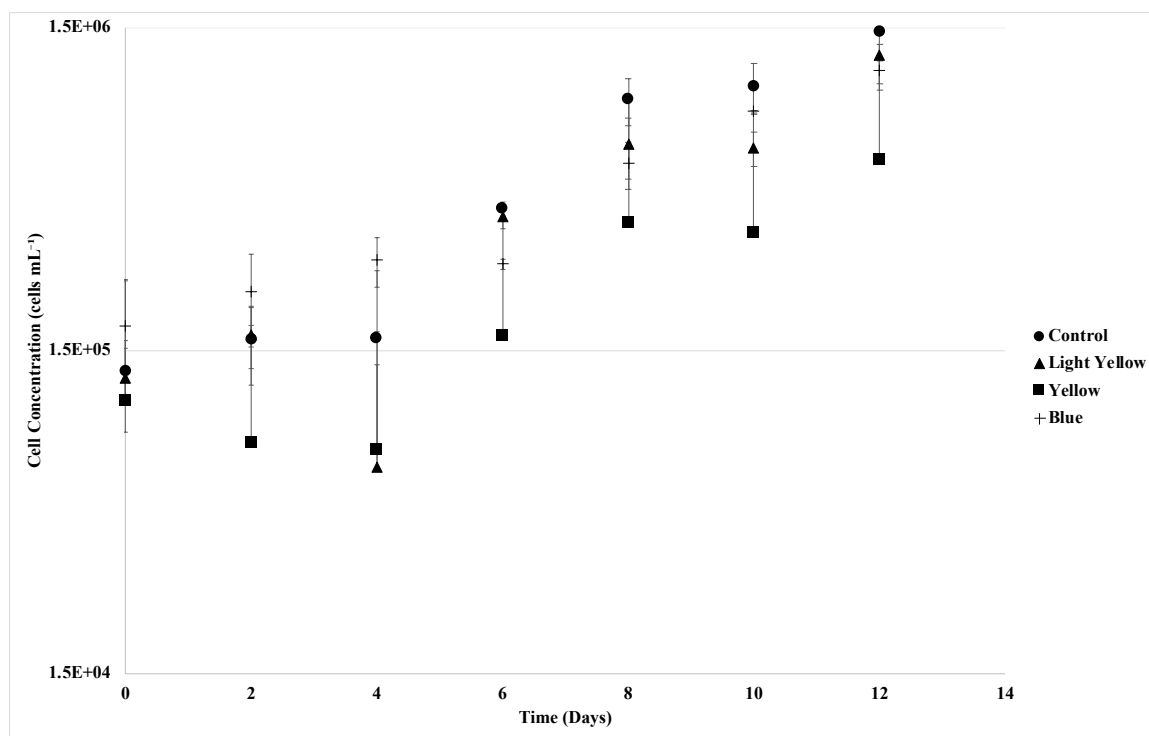


Figure C.7 Cell concentrations for the 4 blue light conditions tested, the no light filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.

Table C.2 Growth conditions (filter types) and doubling times for blue light growth studies. Each condition was grown in duplicate. One replicate in the yellow condition was excluded due to severe clumping.

Filter	Final Cell Count	Doubling Time (h)
	(cells mL ⁻¹)	
	Average ± Standard Deviation	Average ± Standard Deviation
Control	1.45e ⁶ ± 1.41e ⁴	38.49 ± 10.55
Light Yellow (#313)	1.23e ⁶ ± 2.62e ⁵	27.04 ± 12.12
Yellow (#6)	1.11e ⁶	40.23 ± N/A
Blue (#384)	1.10e ⁶ ± 9.19e ⁴	61.45 ± 2.66

The highest observed chlorophyll concentrations were in the blue (#384) filtered cultures (Figure C.8) correlated with the lowest PAR. However, the filtered RGd-1 cultures were similar in chlorophyll concentration to the no-filter control. The chlorophyll enzyme, glutamate-1-semialdehyde aminotransferase (gsa) is blue light-induced in *Chlamydomonas reinhardtii*.²⁶⁵

This enzyme catalyzes pyridoxal 5'-a phosphate-dependent reaction which converts glutamate-1-semialdehyde (GSA) to δ -aminolevulinate (ALA), which is the first committed step in porphyrin biosynthesis.²⁶⁶ The decreased PAR in the blue (#384) filtered cultures may have resulted in increased chlorophyll concentrations. The low PAR, high chlorophyll concentration has been observed in other algae cultures.^{267, 268}

The two culture conditions resulting in the highest final Nile Red fluorescence 302.5 ± 13.4 and 231 (one tube) also had the highest blue light intensities, blue (#384) and yellow (#6) (Figure C.9). However, this trend did not remain consistent for the two lower blue light intensities, light yellow (#313), and the control. The blue (#384) culture condition grew the slowest of the four conditions with a 61.45 h doubling time (Table C.1, Figure C.3). Increased TAG under higher blue light conditions can be explained by increased activity of carbonic anhydrase and ribulose biphosphate carboxylase/oxygenase (Rubisco), both of which have greater activity under blue light.^{269, 270} It is important to recognize that the cultures in this study never reached maximum Nile Red fluorescence due to clumping in some of the tubes. However, trends can still be observed. There may be different optimal blue light intensities for TAG and for doubling time. Future work is required to elucidate this effect.

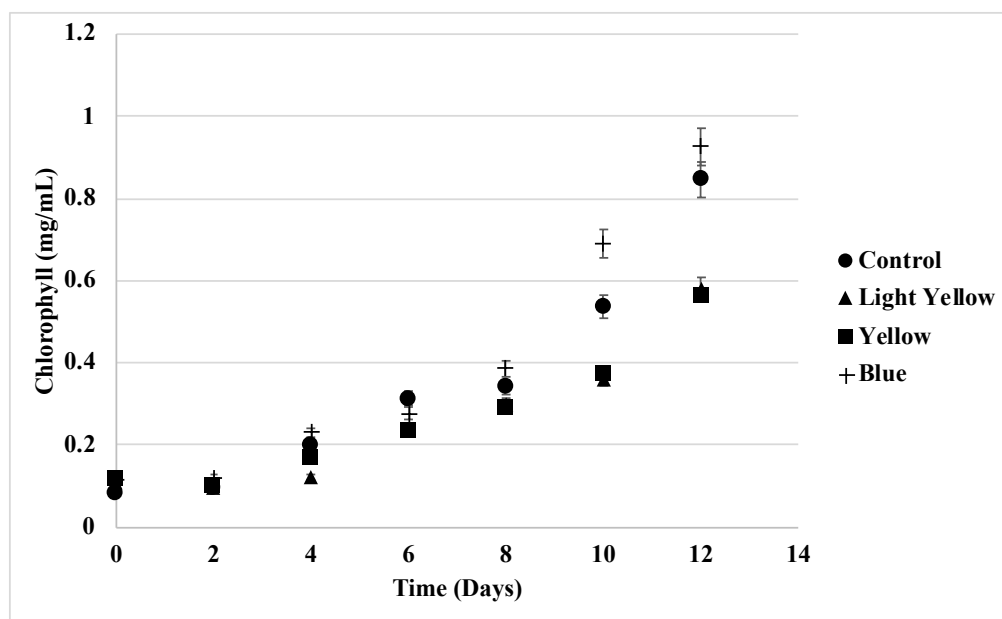


Figure C.8 Total chlorophyll concentrations (mg/mL) for each of the 4 blue light conditions tested, the no-filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.

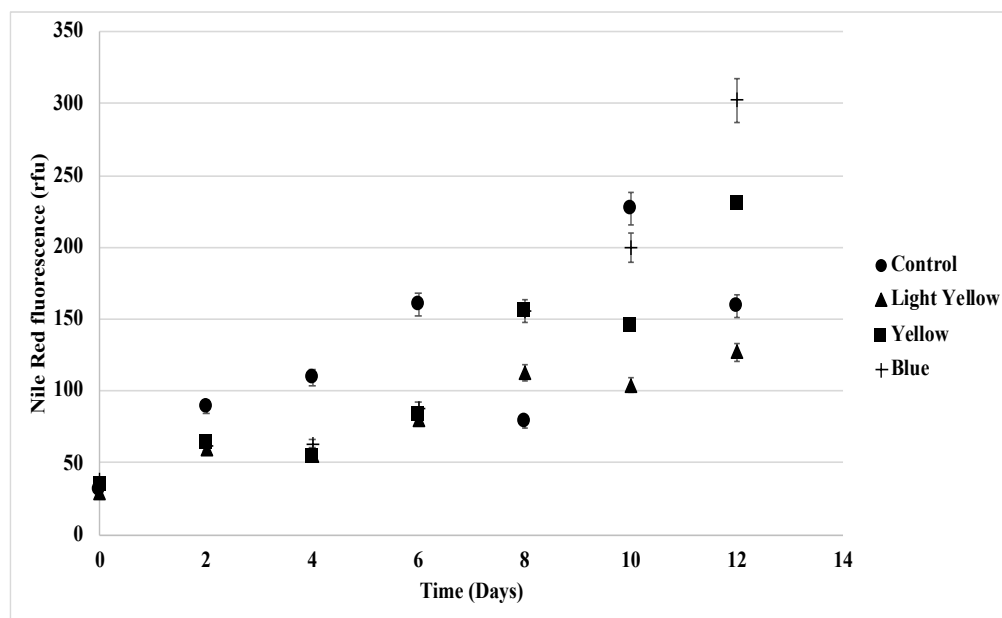


Figure C.9 The Nile Red fluorescence (rfu) for the 4 blue light conditions tested, the no-filter control, light yellow (#313), yellow (#6) and blue (#384). The error bars represent the standard deviation of the mean.

Conclusions

It was expected that there would be a clear effect of blue light supplementation on RGd-1 growth, cellular morphology, and lipid productivity. In particular, it was thought there would be a decreased doubling time with increased blue light intensity. RGd-1 cultures grown with the light yellow filter (#313) were consistent with this trend where the highest blue light intensity resulted in the lowest doubling time compared to the other growth conditions (Table C.1). The highest total Nile Red fluorescence was found in the yellow filter condition (#6). The blue filter (#384) condition resulted in the highest doubling time and highest Nile Red fluorescence. However, the overall PAR (Table C.1) was considerably lower compared to the other conditions resulting in a slower doubling time (Table C.2), which resulted in higher concentrations of lipids. These results indicate that the light yellow filter may be promising as an additional strategy for decreasing the RGd-1 growth rate.

Algal biofuels is an emerging technology, but further research is required to decrease production costs. Diatom strain, RGd-1, is a very promising candidate for large-scale algal biofuel applications due to its high TAG and biodiesel content at 30-40% (w/w) and 70-80% (w/w), respectively. To increase the growth rate, other low-cost strategies are required, such as supplementing cultures with blue light. Understanding the genes expressed under the different blue light conditions will elucidate the role of blue light on the cell cycle which may be highly correlated with the growth rate. Further, it is important to perform fundamental studies of diatom growth that provide the best opportunities to maximize understanding of unique aspects of diatom physiology.

To understand the effects of different blue light intensities on RGd-1 growth, it is important to determine the changes in global gene expression. Given the light-dependent

activation of aureochromes that regulate dsCYC2 at the G1/S phase of the cell cycle, it is expected that there would be increased expression for gene encoding cyclin D, which is prominent during the G1 phase of the cell cycle.²⁸ Additionally, diatom specific cell cyclins dsCYC1, dsCYC2, dsCYC5, dsCYC6, dsCYC7, dsCYC8, dsCYC9, and dsCYC11 mRNA were found to be more prevalent during G1 or S phases for *P. tricornutum*,²⁸ and we expect to see increased expression at G1/S for RGd-1.

APPENDIX D

THE EFFECTS OF ARSENIC SUPPLEMENTATION ON RGd-1 GROWTH RATE AND
LIPID ACCUMULATION

Introduction

Green algae and diatoms have the ability to take up and transform arsenic in a variety of aquatic environments. Microbial bioremediation, including algae, may provide a more cost-effective and environmentally less detrimental way of remediating heavy metals such as arsenic.²⁷¹ Arsenate and arsenite can be taken up through the cell via phosphate transporters and aquaglyceroporins, respectively.^{272, 273} The RGd-1 culture is especially interesting because arsenic resistance and reductase genes have been identified in both the diatom and an associated bacterium, *Brevundimonas* sp. with an assembled genome.

Arsenic resistance in bacteria is under the control of the ars operon.^{272, 274, 275} The ars operon contains three genes, arsR (trans-acting repressor), arsB (membrane-bound arsenite permease pump) and arsC (intracellular arsenate reductase).²⁷² ArsR senses the presence of As(III) and controls the expression of arsB and arsC. ArsC reduces arsenate to arsenite using glutathione as a reducing agent. ArsB functions as an $\text{As}(\text{OH})_3^- \text{H}^+$ antiporter resulting in the As(III) out of the cell. In some bacteria such as *E. coli*, additional genes are co-located in the ars operon, arsD, and arsA. *E. coli* has also been found to have two efflux proteins, ArsB, and ArsAB that facilitate the removal of arsenite from the cell.

Arsenate (AsO_4^{3-}) is chemically similar to phosphate (PO_4^{3-}) and is, therefore, a competitive inhibitor of phosphate uptake at high concentrations.²⁷³ Decreased phosphate concentrations for some microalgae cultures have led to an increase in arsenate uptake.²⁷⁶ Arsenate uptake could be reduced by increasing the PO_4^{3-} concentrations in *Chlorella salina* and *Skeletonema costatum*,²⁷⁷ and in *Chlamydomonas reinhardtii*.²⁷² Arsenate reduction has been identified in *Chlorella*, *Skeletonema costatum*, and *Thalassiosira*.^{276, 278} Previously, *Chlorella* has been found to convert

arsenate to dimethylarsenic species.²⁷⁸ Interestingly, this *Chlorella* was isolated from hot springs in Japan, which contain higher levels of arsenic compared to other freshwater areas such as non-volcanic or non-hot spring streams and lakes.²⁷⁸ The cyanobacteria, *Synechocystis* and Cyanidiales can oxidize As (III) to As (IV).^{273, 279}

Arsenic is introduced into the environment by natural occurrences such as weathering of rocks, volcanic or hot spring activity, or by anthropogenic activities such as pesticides, mining, the combustion of fossil fuels or wood preservatives.^{272, 280} In marine environments, algae and other phytoplankton that are capable of arsenic uptake, convert it into dimethylarsenic species, which exists primarily as arsenosugars catalyzed by the enzyme, arsenite methyltransferase (ArsM). When consumed up the food chain, the dimethylarsenic species are converted to trimethyl arsenic species (arsenobetaine). The arsenosugars and arsenobetaine that bioaccumulate up the food chain have much lower toxicity compared to arsenate/arsenite for mammals.²⁷⁸ Arsenoribosides have been detected in *Chaetoceros concavicornis*, *Chlorella vulgaris*, *Monoraphidium arcuatum*, *Chlamydomonas*, *Dunaliella*, *Phaeodactylum*, *Thalassiosira*, and the cyanobacteria, *Synechocystis* and *Nostoc*²⁷². Lipid-soluble arsenic compounds have been detected in *C. vulgaris*, *C. ovalis*, *C. pyrenoidosa*, *D. tertiolecta*, *P. tricornutum*, *S. costatum*, and *T. pseudonana* and the cyanobacteria, *Oscillatoria rubescens* and *Synechocystis*²⁷². The major arsenic species in freshwater systems are inorganic arsenic and methylarsenicals including dimethyl- and trimethylarsenic species.²⁷⁸

Other diatoms besides RGd-1 have been found in arsenic-contaminated water. *Skeletonema costatum* was found to reduce methylated arsenic (III) at NW Netley Buoy and Calshot Buoy near Hampshire, England.²⁸¹ Methylation and reduction of arsenate and arsenite have been found primarily in photic zones and taken up by phytoplankton, including algae. Further,

methylarsenicals and arsenite are have been located in freshwater systems with concentrations similar to marine concentrations.²⁸² Witch Creek has been found to have approximately 300 ppb arsenic, or thirty times the upper limit of the EPA allowable drinking water standard of 10 ppb.^{283,}

284

Not all diatoms can utilize arsenic. The main effect on *Achnanthes minutissimum* was a reduction in cell size during acute exposures.²⁸⁵ Arsenic has been found to inhibit growth and oxidative phosphorylation in phytoplankton.^{277, 286, 287} Specifically, arsenic uptake by diatoms has been found to occur when phosphate concentrations are low.²⁷⁷ Diatoms were shown to reduce arsenate to dimethylarsenic acid in aquatic environments, which results in a stable, less toxic arsenic form.

While arsenic cannot be degraded from contaminated sites, it can be transformed to a less toxic state. Several methods of arsenic remediation have been identified.²⁷³

1. Biotransformation using arsM. Many bacteria can convert arsenate to volatile methylated species in aqueous or sub-surface soil systems.²⁷³
2. Sulfate reduction. With sulfate as the electron acceptor, sulfate-reducing bacteria (SRB) have been found to generate sulfide which can then react with arsenic. The sulfide-arsenic complex has low solubility causing the complex to precipitate. Arsenic can be precipitated as arsenates, arseno-sulfides or co-precipitated with other metals.²⁷³
3. Ferrous oxide removal. When As (III) is oxidized to As (V), it becomes adsorbed to Fe(III) and is removed from aqueous systems or immobilization in soil systems.²⁷³

According to Figure D.1, the arsenic was predominantly in the form of HAsO_4^{2-} at pH 9.3 when RGd-1 was collected from Witch Creek. Inorganic forms of arsenic such as arsenate, As

(V), and arsenite, As (III), are the most toxic. As (V) can become incorporated into phosphorylated compounds that will be used for ATP synthase or by blocking (sulfhydryl) SH groups which would interfere with enzymes such as pyruvate dehydrogenase or 2-oxoglutarate dehydrogenase resulting in membrane leakage and cell death due to the production of reactive oxygen species.²⁸⁸

For diatom strain, RGd-1, early results had shown a 28% decrease in doubling time when grown artificial Witch Creek water (Bold's Basal Medium with added Witch Creek chemicals) compared to when sodium arsenate was not included in the artificial Witch Creek water. Similarly, a decrease in doubling time was found when RGd-1 was grown in the presence of sodium arsenite (25 ppb) compared to the control without arsenite. Both datasets indicated that there was potential for increased TAG and decreased doubling time in the presence of sodium arsenate in modified Bold's Basal Medium (B8.7SiS – Bold's + 2 mM sodium metasilicate and ASP2 concentrations of B12 and other vitamins) as a diatom growth medium.^{49, 117, 118} It has been suggested that some microbiota detoxify arsenic by attaching the metalloid to lipids as arsenolipids.²⁸⁹ Further, some algae have been found to hyper-accumulate and transform arsenic from the surrounding water.²⁷²

Due to these initial testing results, it was hypothesized that the presence of arsenic might decrease the RGd-1 doubling time and that arsenic would be stored as arsenolipids that can be converted for use in biofuels on the high-lipid producing diatom, RGd-1. Here, both arsenate and arsenite were added to B8.7SiS and different molar ratios of phosphorus to arsenic were tested to determine the optimal ratio for RGd-1 growth.

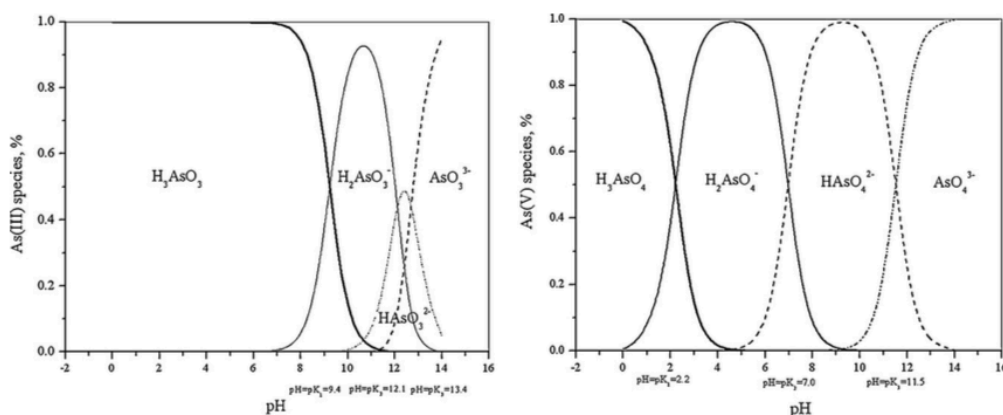


Figure D.1 Speciation of As(III) and As(IV) across pH ranges -2 to 14 in water.^{57, 58}

Methods

There were two sets of experiments that exposed RGd-1 to arsenic; (1) initial testing with RGd-1 growing in Witch Creek Water or Artificial Witch Creek Water and (2) growth in sodium arsenate in B8.7SiS.

Initial Testing

The initial testing was performed in triplicate 250 mL flasks. RGd-1 Cultures were grown under 14:10 light/dark (L/D) cycle, aerated with ambient air temperature-controlled incubator at 30°C. The light intensity of the incubator was $\mu\text{mole photons m}^{-2}\text{s}^{-1}$ using twelve T5 four ft fluorescent lights in a square-wave 14:10 light/dark (L/D) cycle. Samples were collected daily just prior to the end of the light cycle. To quantify the growth, a minimum of 100 cells were counted from each sample using a hemacytometer (Reichert). The sample pH was measured using a standard benchtop pH meter (Accumet). The lipid accumulation was measured daily by staining the cultures with Nile Red (0.25 mg/mL suspended in 20% DMSO) and using a

microplate reader (BioTek Synergy). Final dry cell weights (DCW) were assessed at the end of the experiment by filtering 25 mL of each culture with pre-weighed F/F Glass Microfiber Filters (Whatman). Samples were re-weighed after drying at 60°C after approximately 24 and 48 hours.

RGd-1 cultures were grown in water collected from Witch Creek (Witch Creek Water) and Artificial Witch Creek Medium. Because Witch Creek water is in limited supply, the chemicals from Witch Creek that were not present in B8.7SiS were added to create an artificial medium named Artificial Witch Creek Medium ((AWCM) Table D.1) used in later growth studies. Based on ion chromatography (Dionex ICS-1100 Ion Chromatography System), inductively coupled plasma-mass spectrometry (ICP-MS (Agilent Technologies 7500ce) and nitrogen analysis, it was possible to determine the major chemicals constituting Witch Creek water. The AWCM, therefore, contained all measured Witch Creek chemicals, except antimony.

Ash Free Dry Weight (AFDW)

The ash-free dry cell weight was determined by heating the DCW sample and filter to 500°C for five hours. The samples were weighed immediately after removing from the furnace. The difference between the remaining sample DCW and filter was the ash-free dry weight.^{49, 290,}

Table D.1 2013 Witch Creek water analyses and B8.7SiS chemical concentrations.

	Creek Water	Bold's Basal Medium
Sodium (ppm)	113.065	114.719
Magnesium (ppm)	13.129	7.389
Aluminum (ppb)	0.069	0.000
Silicon (ppm)	72.105	56.170
Potassium (ppm)	8.320	82.779
Calcium (ppm)	7.200	6.813
Vanadium (ppb)	0.001	0.000
Manganese (ppm)	0.782	0.399
Iron (ppm)	3.490	1.000
Cobalt (ppb)	0.349	0.099
Copper (ppb)	0.872	0.399
Zinc (ppb)	0.149	2.001
Arsenic (ppb)	0.300	0.000
Selenium (ppb)	0.004	0.000
Molybdenum (ppb)	0.095	0.473
Cadmium (ppb)	0.001	0.000
Antimony (ppb)	0.011	0.000
Barium (ppb)	0.010	0.000
Lead (ppb)	0.002	0.000
F (ppm)	130.820	0.000
Cl (ppm)	709.960	27.952
SO ₄ ²⁻ (ppm)	510.260	34.468
TC (ppm)	31.130	18.976
DIC (ppm)	13.525	0.000
TOC (ppm)	1.995	18.976
TN (ppm)	0.125	0.000

FAME analysis using GC-MS

Harvested biomass was lyophilized to remove all the water content. The biomass (20 - 30 mg) was added to borosilicate culture tubes (Pyrex) and capped with Teflon caps. To the biomass, 1 mL of toluene and 2 mL of sodium methoxide were added and the mixture heated at 90°C for 30 minutes while vortexing every 10 minutes (Fischer Scientific). Two mL of 14% boron trifluoride

(14%) in methanol (Sigma Aldrich) were added to room temperature-cooled samples and heated for an additional 30 minutes with vortexing every 10 minutes. Following this treatment, the samples were cooled and 0.8 mL of hexanes (Fischer Scientific) and 0.8 mL of saturated NaCl solution were added to facilitate separation of FAMES into the organic phase. Samples were reheated to 90°C for 10 minutes to further facilitate phase separation and then centrifuged for two minutes at 6 ×g. The top organic layer was extracted using a glass syringe (Hamilton) and then run on gas chromatography-mass spectrometry (GC-MS) (Agilent) using a DB-23 column (Restek) to quantify FAMES against NLEA FAME mix (Restek).

Determination of the optimal P:As ratio

Cultures grown with increased phosphate concentrations have been found to take up less arsenate. Previous studies with *Agrobacterium tumefaciens* have found an optimal P:As ratio of 10:1.²⁹² RGd-1 was grown in 5 different phosphorus to arsenic ratios to determine what would be the optimal parameters for growth; control (B8.7SiS P:As ratio, 1:1, 2:1, 5:1 and 10:1) (Table D.2).

Table D.2 Molar Phosphorus and arsenic ratios used in arsenate experiments.⁶¹

P:As	Control	10:1	5:1	2:1	1:1
P	1.63 mM	40 µM	20 µM	8 µM	4 µM
As	0	1 µM	1 µM	1 µM	1 µM

The 5:1 P:As ratio was selected to carry out the remainder of the experiments because it resulted in the highest cell count and was among the fastest doubling times. Therefore, all other experiments maintained the 20 μM P concentration while varying the As concentration in the growth medium. Low and high sodium arsenate ranges were targeted (low = 10-90 ppb As; high = 100-900) (Table D.3).

Table D.3 Descriptions of arsenate experiments. Seven experiments were performed with different concentrations of sodium arsenate. The first set of experiments was used to determine the optimal As:P ratio for RGd-1. That phosphorus concentration was used for all future experiments with varying As.

Experiment	Description
1	P:As Ratios (10:1, 5:1, 2:1, 1:1; 0 As)
2	High As concentrations (0, 150, 300, 600, 900 ppb) - Discounted due to undefined period of time with lights on in Cobleigh 307 (Days 14-16). P=20 μM
3	Low As concentrations (0, 10, 25, 50, 75 ppb). P=20 μM
4	High As Concentrations - repeated (0, 150, 300, 600 ppb). P=20 μM
5	Filling in low As concentrations (0, 30, 40, 60, 90 ppb). P=20 μM
6	Bold's Basal Medium PO_4^{3-} concentration = 1.63 mM (As concentrations = 0, 10, 50, 100, 500)
7	0 P (As concentrations = 0, 10, 50, 100, 500)

Results

A consistent arsenic concentration of 262.5 ppb has been measured over multiple years since 2009. It was hypothesized that RGd-1 may have adapted to survive and succeed in high arsenic concentrations. Specifically, (1) the lack of arsenic in the RGd-1 growth medium may contribute to a changeable growth rate and cellular morphology (2) RGd-1 assimilates arsenic into arsenolipids or arsenosugars (3) RGd-1 respire arsenic reducing arsenate to arsenite. The following series of experiments were designed to determine the effect of arsenic on RGd-1

growth and lipid accumulation. RGd-1 was grown in two sets of experiments (1) (initial testing in Witch Creek Water or AWCM) and (2) B8.7SiS + sodium arsenate.

Initial testing

The fastest doubling times occurred when RGd-1 was grown in Witch Creek water with Bold's additions. Witch Creek water is low in phosphate and nitrate (Table D.1). However, this growth condition resulted in the lowest Nile Red fluorescence of all of the conditions tested at that time (Figure D.2, Table D.3). The slowest doubling times also occurred when twice the amount of iron was added to the AWCM (Figure D.3, Table D.4). There was one condition that resulted in no growth, Witch Creek Water with Bold's additions titrated to pH 8 with Bold's concentration of ethylenediaminetetraacetic acid (EDTA).

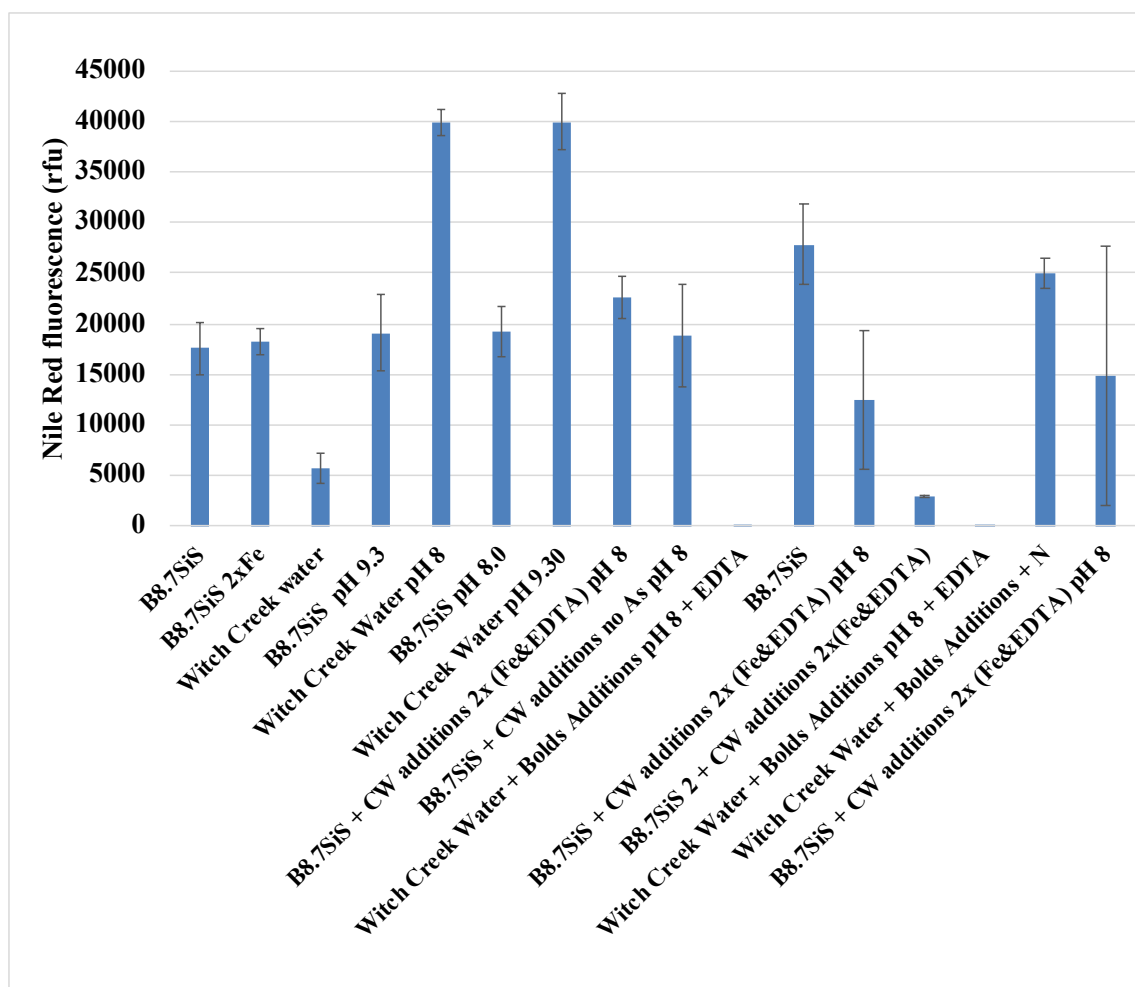


Figure D.2 The Nile Red fluorescence for the initial testing with sodium arsenate. The error bars represent the standard deviation of the mean.

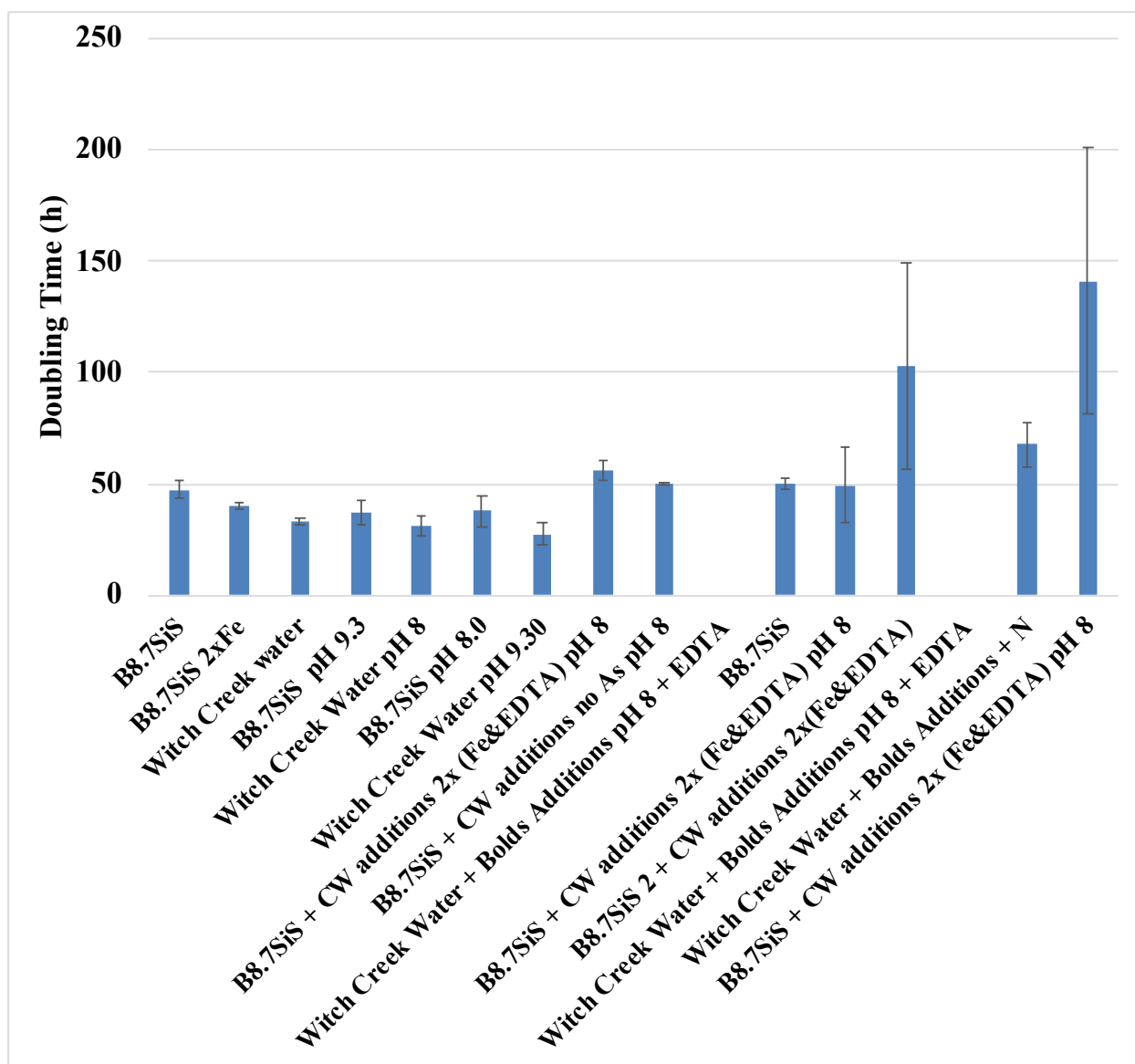


Figure D.3 The doubling times for the initial testing with sodium arsenate. Two conditions did not grow; Witch Creek Water with Bold's additions at pH 8 + EDTA. This condition was repeated to verify the result. The error bars represent the standard deviation of the mean.

Table D.4 Doubling times and DCW for the RGd-1 initial testing with sodium arsenate.

Samples	Average doubling time (h)	DCW (g/L)
Rgd1 Growth rate experiment-2		
B8.7SiS	57.45	0.536
Witch Creek Water	53.67	0.139
Bold's Media, 2xFe, pH 8.70, [Fe]/[EDTA]=0.30	55.81	0.424
Rgd1 Growth rate experiment-3		
Witch Creek Water w/ Bold's Media Additions pH 8.00, 0.51 M Nitrogen, [Fe]/[EDTA]=0.47	30.72	0.503
Witch Creek Water w/ Bold's Media Additions pH 9.30, 0.51 M Nitrogen [Fe]/[EDTA]=0.47	36.00	0.543
Bold's Media w/ Witch Creek Water Additions pH 8.00, [Fe]/[EDTA]=0.47	37.02	0.432
Bold's Media w/ Witch Creek Water Additions pH 9.30, [Fe]/[EDTA]=0.47	38.53	0.463
Growth rate experiment-4		
Bold's Media, pH 8.70	51.92	0.508
Bold's Media w/WC additions, no As, pH 9.30, [Fe]/[EDTA]=0.47	78.90	0.488
WC w/Bold's Media additions, 2x EDTA, pH 9.30, [Fe]/[EDTA]=0.15	no growth	0.187
Bold's Media w/WC additions, 2x Fe&EDTA, pH 9.30, [Fe]/[EDTA]=0.30	39.59	0.336
Bold's Media, 2x Fe&EDTA, pH 9.30, [Fe]/[EDTA]=0.15	52.05	0.491
Growth rate experiment-5		
Bold's Media, 2x Fe&EDTA, pH 9.30, [Fe]/[EDTA]=0.15	148.08	0.153
Witch Creek Water w/Bold's Media additions, 2x EDTA, pH 9.30, [Fe]/[EDTA]=0.15	no growth	0.024
Witch Creek Water w/ Bold's Media Additions, 2.97 M Nitrogen (right amount) pH 9.30, [Fe]/[EDTA]=0.47	68.81	0.480
Bold's Media w/WC additions, 2x Fe&EDTA, pH 9.30, [Fe]/[EDTA]=0.30	91.99	0.360
K.M. Moll et al 2014 (2 mM Si) (tube reactors)	28.28	0.500

Arsenate

When grown in different P:As ratios, it was observed that there was a significant increase in cell numbers in the 5:1 P:As condition (Figure D.4). The doubling time was the lowest in the 10:1 and 5:1 conditions. With the increase in cell numbers and low doubling time for the 5:1

condition, it was chosen to proceed forward for the remainder of the studies. The 1:1 condition was the most similar in cell numbers and in doubling time when compared to the B8.7SiS control (Table D.4).

The pH was similar in the 5 conditions tested. There was a significant increase in the 5:1 and 10:1 condition (Figure D.5). However, the 5:1 and 10:1 ratio conditions were within error of each other. According to Figure D.6, there was no significant difference in Nile Red fluorescence between the five conditions.

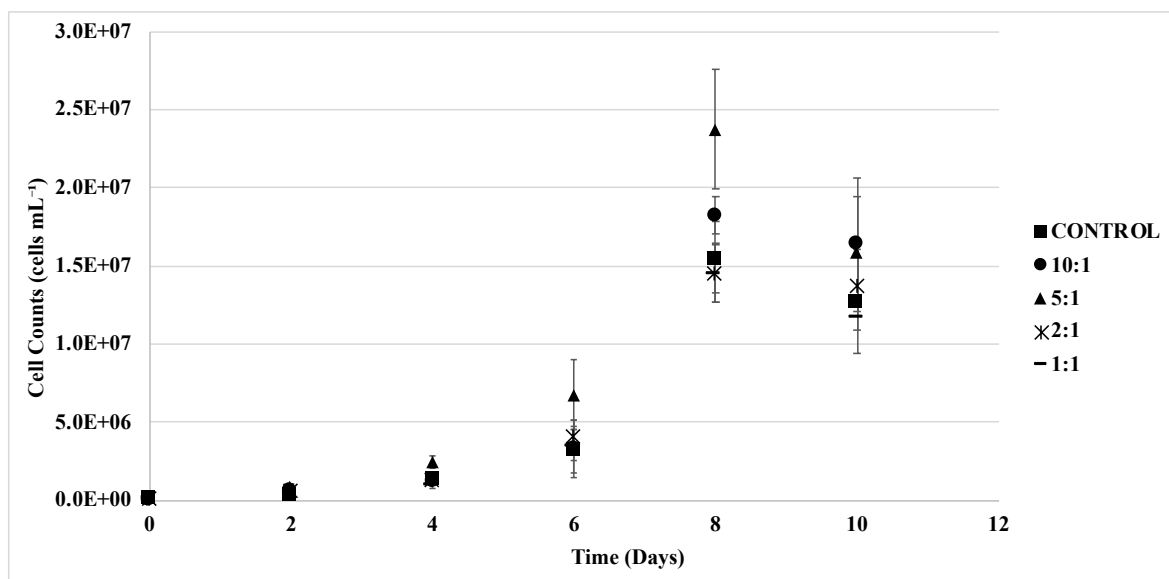


Figure D.4 Cell counts for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.

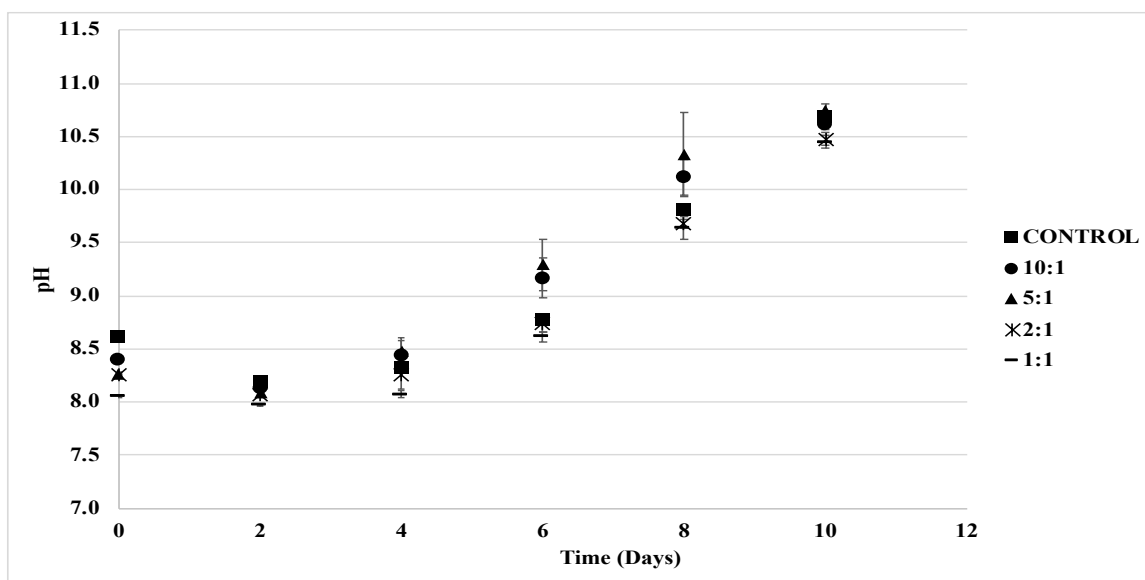


Figure D.5 The pH for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.

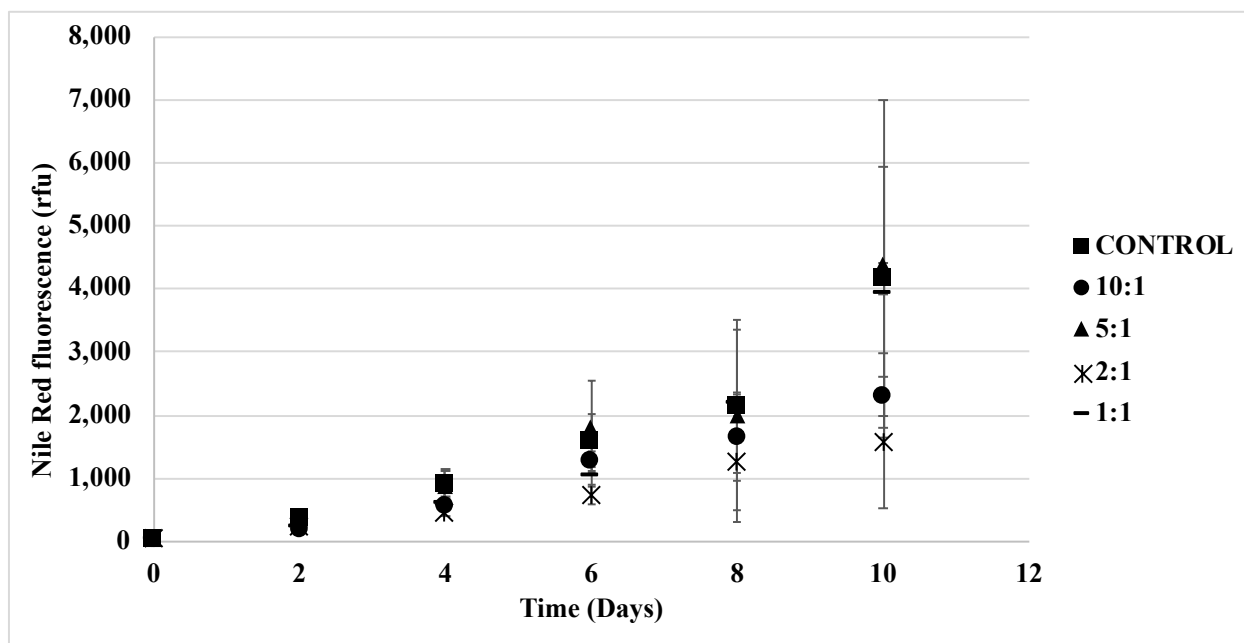


Figure D.6 The total Nile Red fluorescence for four phosphorus to arsenic ratios (10:1, 5:1, 2:1 and 1:1) and a control (B8.7SiS phosphorus concentrations). The error bars represent the standard deviation of the mean.

The fastest doubling times occurred for the 0 phosphorus, 10 and 50 ppb arsenate conditions. These conditions grew significantly faster than the same arsenate concentrations grown at the 5:1 phosphorus to arsenic concentration (Figure D.7). Neither the arsenate nor the phosphorus concentrations affected the AFDW (Figure D.8).

As shown in Figure D.11, the highest % FAMES occurred for cultures grown in the highest arsenate conditions, whereas the lowest % FAMES were found in cultures grown in 0 phosphate conditions (Figure D.11). The percent TAG decreased substantially compared to previous reports of 70-80% FAME.⁴⁹ However, there was substantially lower phosphate in these experiments compared to previous high FAME-accumulating experiments (B8.7SiS concentrations of P (1.63 mM)).¹¹⁷

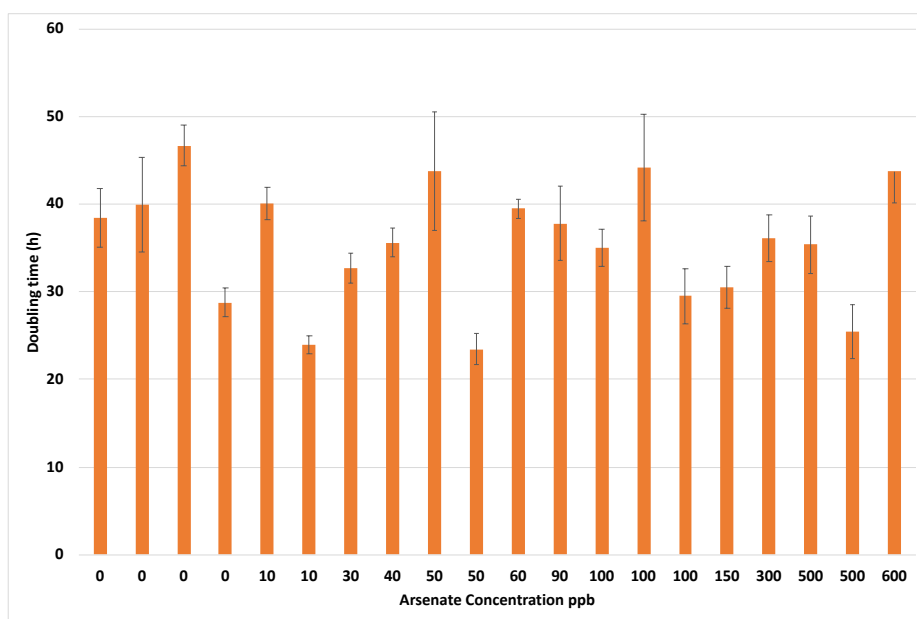


Figure D.7 The doubling times for the different arsenate concentrations tested. The error bars represent the variance resulting from an ANOVA (2-factor without replacement) analysis. The error bars represent the standard deviation of the mean.

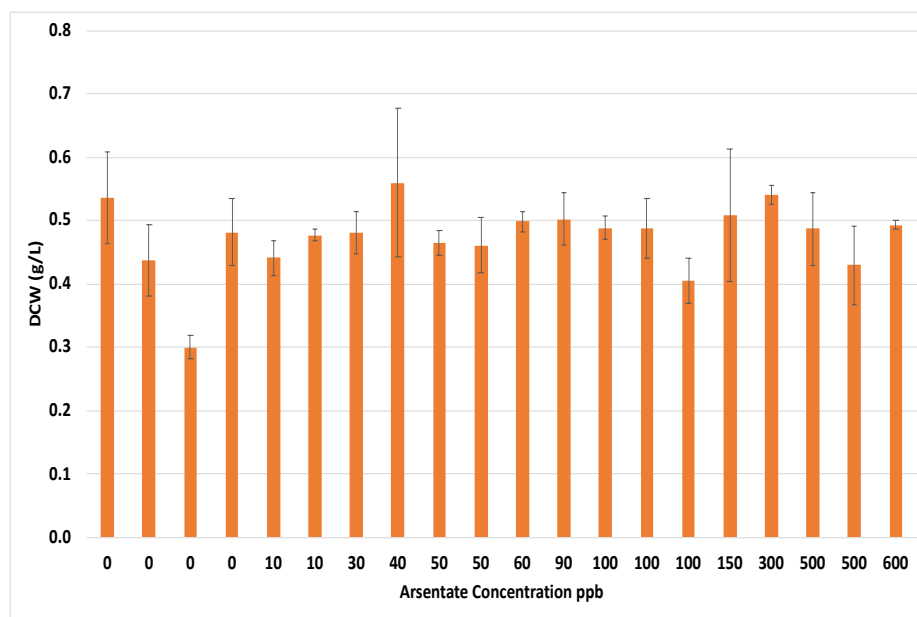


Figure D.8 The average ash-free dry weights for the different arsenate concentrations tested. The error bars represent the variance resulting from an ANOVA (2-factor without replacement) analysis. The error bars represent the standard deviation of the mean.

Discussion

The initial testing revealed a 28% decrease in doubling time when arsenate was present in AWC. However, this decrease was not observed when arsenate was added to B8.7SiS. Witch Creek Water contains an extensive list of chemicals that are not present in B8.7SiS. There may be chemicals (e.g. antimony) in AWC that interact synergistically with arsenic that may contribute to an increased growth rate and TAG accumulation.

In the B8.7SiS + As studies, the TAG concentration decreased considerably compared to previous RGd-1 reports in B8.7SiS. This indicates that the addition of arsenic to B8.7SiS had a negative effect on TAG accumulation. However, there was also substantially lower concentrations of phosphorus in these studies compared to standard Bold's phosphorus concentrations, perhaps resulting in insufficient phospholipids to build lipids.²⁹³

Arsenotriacylglycerols arsenic hydrocarbons and arsenic fatty acids have been observed in marine algae including the green alga, *Coccomyxa*. When transesterified, the arsenophosphate group would be cleaved off resulting in fatty acid methyl esters (FAMES).^{289, 294, 295} When transesterified, the head group containing the arsenic will be cleaved off leaving free fatty acids and a pool of soluble arsenic. Another arsenic precipitation step would be required to remove the majority of the arsenic from the arsenic pool. It may be possible that arsenic exists midchain or at the end of the chain in arsenotriacylglycerols and arsenic fatty acids. In that case, it would not be possible to obtain arsenic-free FAMES, since transesterification cleaves the head group. It is not likely that FAMES containing arsenic will be combusted for use as biodiesel or used for nutraceuticals. Another step would be required to separate arsenic-free FAMES from arseno-FAMES for use in biodiesel combustion. As mentioned above, when diatoms methylate arsenic, it becomes less toxic. A more favorable option would be simply to use diatoms or other algae that are able to assimilate arsenic into the biomass to bioremediate arsenic-contaminated sites.

According to the results from the initial testing, RGd-1 and its potential phycosphere bacteria may have adapted to the chemical constitution of Witch Creek so that in the presence of arsenic, RGd-1 may have a reduced growth rate and increased TAG accumulation. The co-habiting bacterium, *Brevundimonas* sp. strain, KM-427, has been found to contain the following genes in the Ars operon, *arsH* (arsenic resistance), *arsM* (As(III)-methyltransferase, *arsR* (arsenate reductase) and *acr3* (arsenic resistance). These genes are involved in arsenic resistance.^{296, 297} *Brevundimonas* sp., strain KM-427, may contribute RGd-1 success in high arsenic conditions.

The observed RGd-1 doubling time was the lowest when grown in 0 phosphorus conditions with 10 and 50 ppb, and the FAME concentrations were the lowest in these conditions as well. Because arsenic is an analog of phosphate, it can potentially replace phosphate in molecules such as arsenolipids.²⁸⁹ Arsenic may inactivate the phosphate active transport system as well as inhibit glucose metabolism.^{277, 298, 299} This may explain why the doubling time was lowest in the lowest arsenate concentrations. There may have been just enough arsenate to improve the growth rate without negatively impacting other aspects of metabolism like glucose metabolism.

Summary & Conclusions

The fastest growth rates occurred in the 0 phosphorus, 10 and 50 ppb arsenate conditions. These arsenate concentrations may be promising for future work, especially in AWC. However, the low FAME results are currently sub-optimal compared to previous results. Future work at higher phosphorus concentrations may result in high FAME concentrations. Further, it is necessary to determine whether arsenic is assimilated or respired. This work has a potential application for arsenic bioremediation through assimilation into diatom biomass into arsenolipids.

Sodium arsenite – Supplementary data

To determine the effects of arsenite on RGd-1, cultures were grown in B8.7SiS with added sodium arsenite in low (10 -50 ppb) and high (100-600 ppb) concentrations. No other modifications to the growth medium were made. Cultures were grown in 1.25 L photobioreactors illuminated with 400 $\mu\text{mole photons m}^{-2}\text{s}^{-1}$ of light and grown at $27^{\circ}\pm 1^{\circ}\text{C}$. When grown in B8.7SiS + sodium arsenate, there was no consistent effect on doubling time (Figure D.9) or Nile Red fluorescence (Figure D.10).

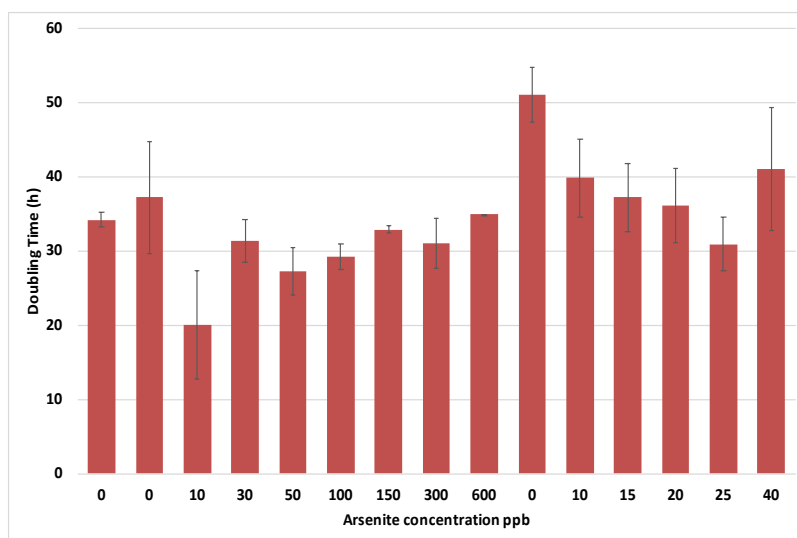


Figure 4 The doubling time for RGd-1 cultures grown in the presence of sodium arsenite in place of sodium arsenate at varying concentrations.

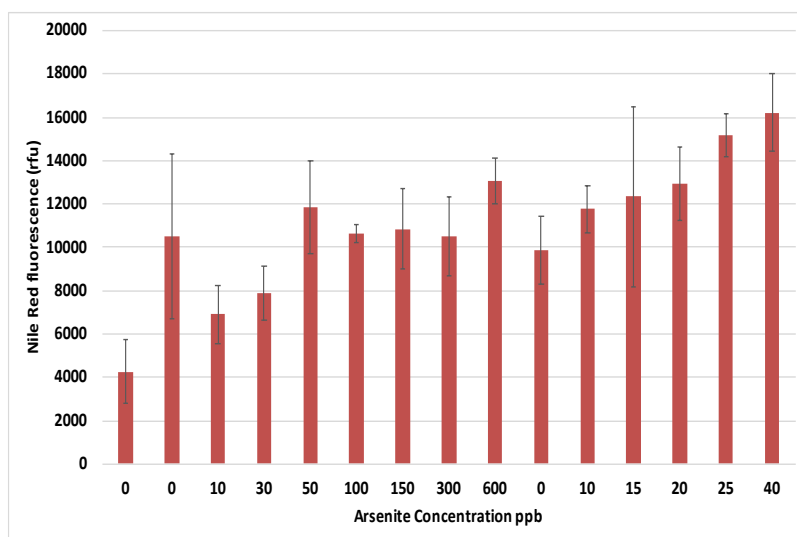


Figure D.10 The Nile Red fluorescence for RGd-1 grown in the presence of sodium arsenite instead of sodium arsenite at varying concentrations.



SEM, Auger and EDX characterization

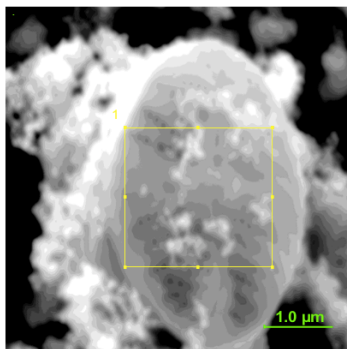


Date : 06/15/2017
 Sample : Diatoms with As
 NanoAuger Report for : Karen Moll
 Prepared by : Manjula Nandasiri



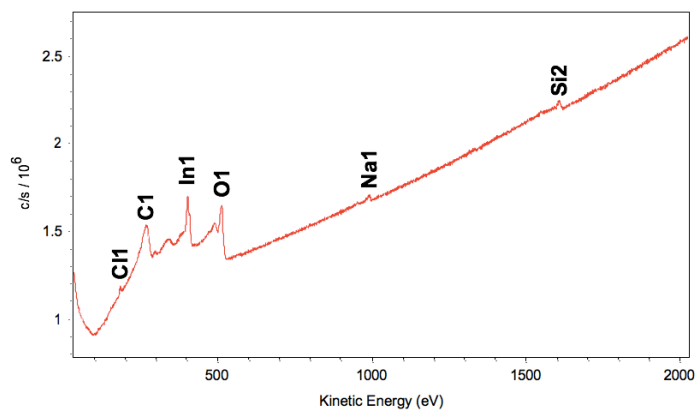
Diatoms sample 1– region 1

SEM image



Auger survey scans

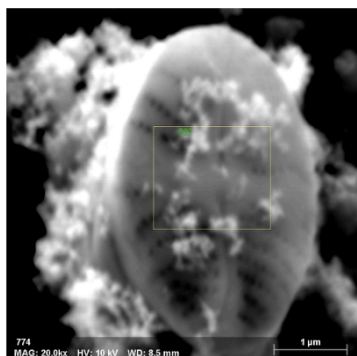
Diatoms As sample1.103.spe



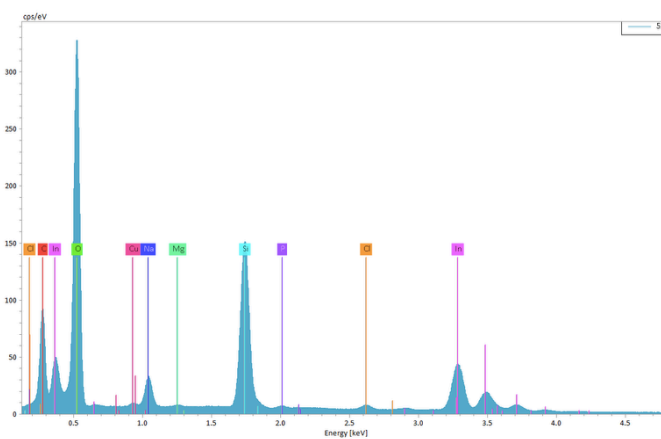
Auger survey scan shows some Si, O, C, Na and Cl. No As was observed.

Diatoms sample 1— region 1

SEM image



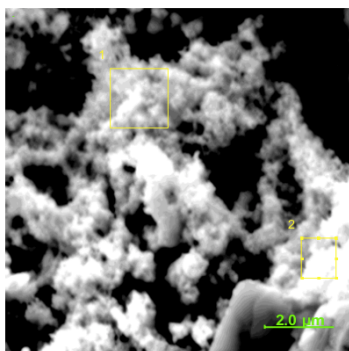
EDX scan



EDX scan shows some Mg, Cu and P in addition to Si, O, C, Na, and Cl observed in Auger scan. No As was observed.

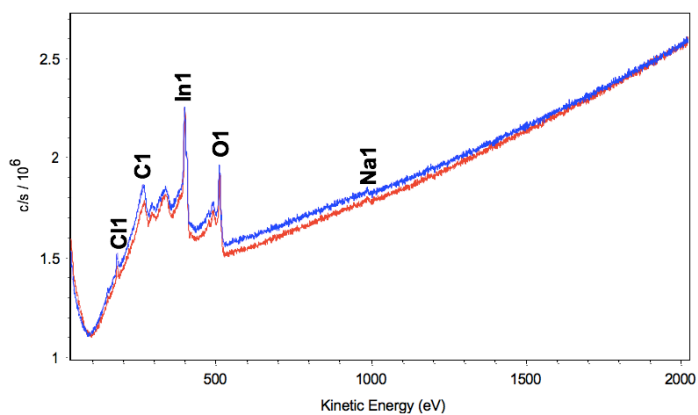
Diatoms sample 1— region 2 (outside the diatom)

SEM image



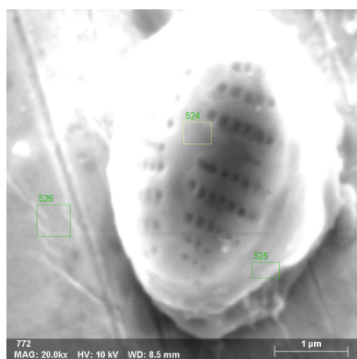
Auger survey scans

Diatoms As sample1.105.spe

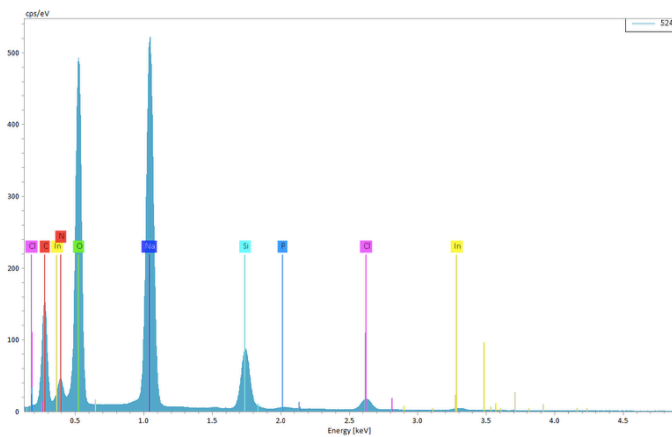


Diatoms sample 1– region 3

SEM image



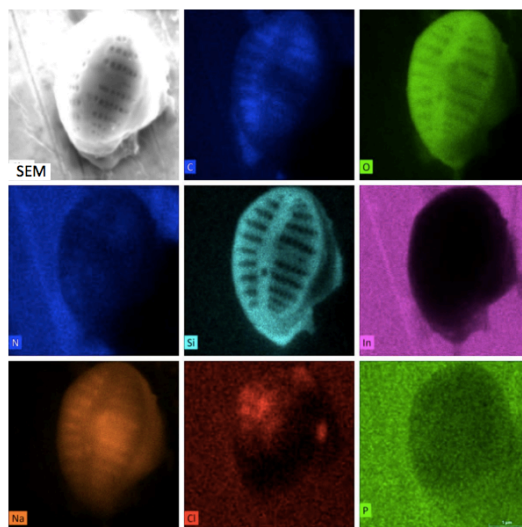
EDX scan



Auger survey scan shows Si, O, C, Na, P and Cl. No As was observed.

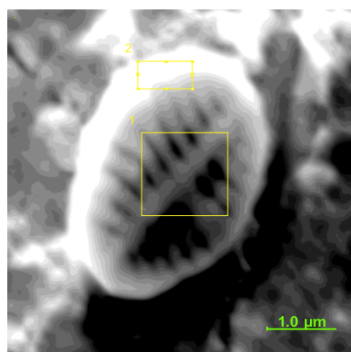
Diatoms sample 1– region 3

EDX elemental maps



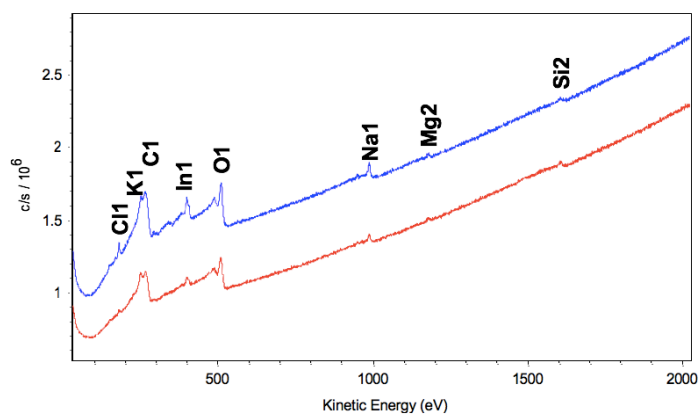
Diatoms sample 2

SEM image



Auger survey scans

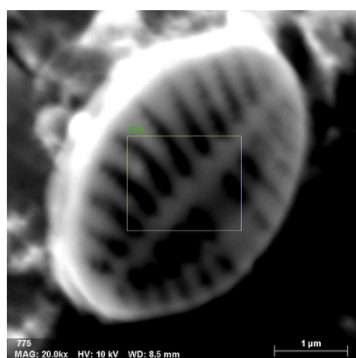
Diatoms As sample2.107.spe



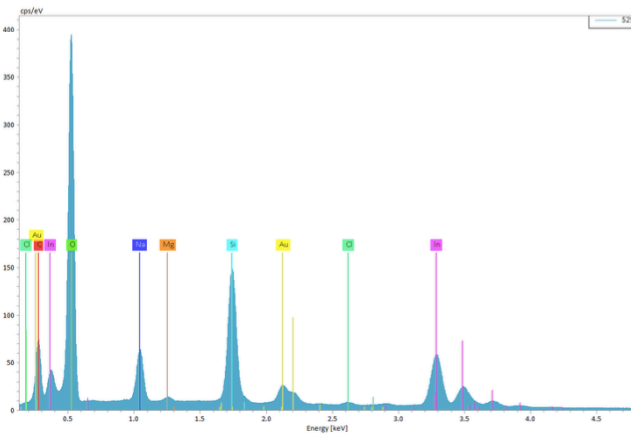
Auger survey scan shows some Si, O, C, Na, Mg, K and Cl. No As was observed.

Diatoms sample 2

SEM image

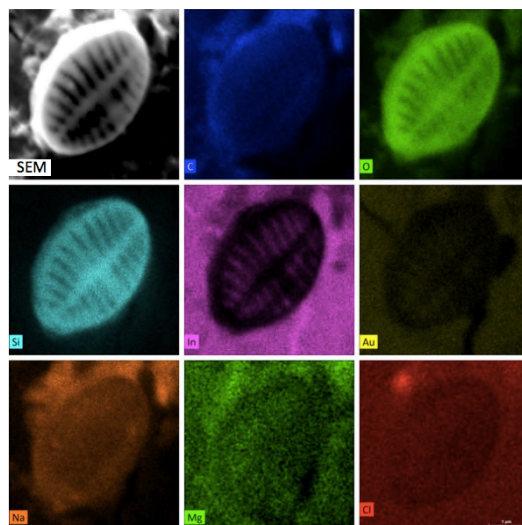


EDX scan



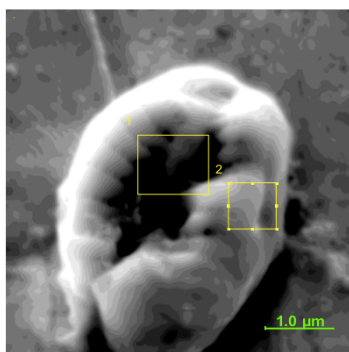
EDX scan shows some Si, O, C, Na, Mg, and Cl on the diatom. In and Au may be coming from the Au coated In foil. No As was observed.

Diatoms sample 2
EDX elemental maps



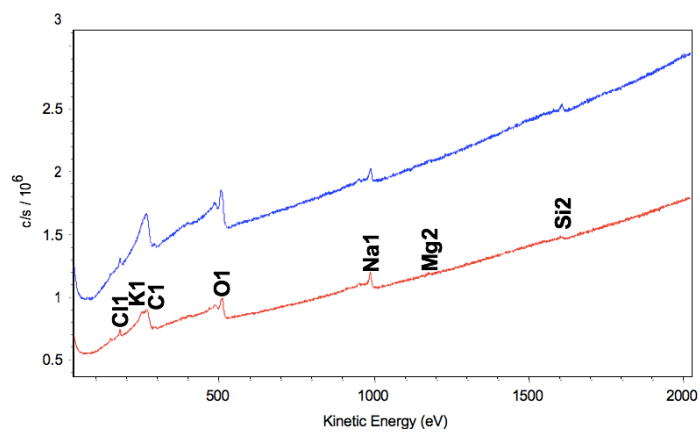
Diatoms sample 3 – region1

SEM image



Auger survey scans

Diatoms As sample3.202.spe



APPENDIX E

STRATEGIES FOR OPTIMIZING BIONANO AND DOVETAIL EXPLORED THROUGH A
SECOND REFERENCE QUALITY ASSEMBLY FOR THE LEGUME MODEL, *MEDICAGO*

TRUNCATULA

Manuscript Information

Karen M. Moll, Peng Zhou, Thiruvarangan Ramaraj, Diego Fajardo, Nicholas P. Devitt, Michael J. Sadowsky, Robert M. Stupar, Peter Tiffin, Jason R. Miller, Nevin D. Young, Kevin A.T. Silverstein, Joann Mudge

BMC Genomics

Status of Manuscript:

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☒ Published in a peer-reviewed journal

18:1

Abstract

Third generation sequencing technologies, with sequencing reads in the tens- of kilo-bases, facilitate genome assembly by spanning ambiguous regions and improving continuity. This has been critical for plant genomes, which are difficult to assemble due to high repeat content, gene family expansions, segmental and tandem duplications, and polyploidy. Recently, high-throughput mapping and scaffolding strategies have further improved continuity. Together, these long-range technologies enable quality draft assemblies of complex genomes in a cost-effective and timely manner. Here, we present high quality genome assemblies of the model legume plant, *Medicago truncatula* (R108) using PacBio, Dovetail Chicago (hereafter, Dovetail) and BioNano technologies. To test these technologies for plant genome assembly, we generated five assemblies using all possible combinations and ordering of these three technologies in the R108 assembly. While the BioNano and Dovetail joins overlapped, they also showed complementary gains in continuity and join numbers. Both technologies spanned repetitive regions that PacBio alone was unable to bridge. Combining technologies, particularly Dovetail followed by BioNano, resulted in notable improvements compared to Dovetail or BioNano alone. A combination of PacBio, Dovetail, and BioNano was used to generate a high quality draft assembly of R108, a *M. truncatula* accession widely used in studies of functional genomics. This strategy proved efficient and cost-effective for developing a quality draft assembly compared to traditional reference assemblies. As a test for the usefulness of the resulting genome sequence, the new R108 assembly was used to pinpoint breakpoints and characterize flanking sequence of a previously identified translocation between chromosomes 4 and 8, identifying more than 22.7 Mb of novel sequence not present in the earlier A17 reference assembly.

Background

Next generation sequencing technologies such as 454, Illumina, and SOLiD became available in the late 2000s.^{17, 300} These technologies have the advantage of extremely high throughput and much lower cost per sequenced base compared to Sanger sequencing.^{26, 301-303} Long read sequencing technologies, such as PacBio and Oxford Nanopore, produce reads in the tens- of kilo-base range, much longer than what was possible even with traditional Sanger technology. However, they also have higher error rates, lower throughput, and higher costs per base compared to the short read technologies. Recently, PacBio throughput and cost per base have improved to the point that *de novo* plant genome assemblies using only PacBio are possible.^{304, 305}

Concomitantly, the throughput and cost of long-range scaffolding and mapping technologies that can increase continuity of an assembly have also improved dramatically. Traditional physical maps, dependent on expensive BAC library preparation, have given way to a variety of new technologies, including Opgen, Keygene, BioNano, and Nabsys maps.^{25, 306-309} BioNano is a high throughput optical mapping technology that utilizes endonucleases to nick long DNA molecules at the enzyme's recognition site, incorporating fluorescent nucleotides to obtain sequence-based patterns. The specific patterns are then used to assemble DNA molecules into a larger genome map, which can then be used to direct and improve a *de novo* genome assembly.³¹⁰

Genomic architecture analyses also can be achieved by sequencing libraries produced from chromatin proximity ligation methods such as Hi-C.³¹¹ Dovetail Chicago libraries are similar to Hi-C but rely on library preparation from *in vitro* rather than *in vivo* reconstituted chromatin that has been cross-linked and sheared. Dovetail Chicago libraries also use extraction

of high molecular weight DNA extraction which limits input DNA length compared to Hi-C, which uses intact chromosomes. These libraries retain proximity signal with sequences physically close together being linked more often than those farther apart. This generates sequence pairs with insert sizes that can be as large as the size of the input DNA, typically ~100kb, for use in scaffolding with Dovetail's in-house software.²⁹

Although BioNano and Dovetail are both long-range scaffolding technologies, there are several important differences. While both rely on restriction endonuclease digestions, different restriction enzymes are used for both technologies, potentially introducing different regional biases. Dovetail and BioNano also differ in the way they handle gaps. Dovetail does not attempt to size the gap, but instead adds 100 Ns between scaffolds that it joins. By contrast, BioNano estimates gap size. Consequently, BioNano can appear to increase scaffold size more when the same scaffolds are joined with both technologies. In addition, BioNano does not automatically split sequences while Dovetail does. BioNano produces a file with possible chimeric sequences, but splitting of these sequences requires manual intervention by the user.

These new sequencing and mapping technologies have increased throughput, driven down costs, and introduced important technological advantages facilitating the sequencing of plant genomes, which are notoriously difficult due to large-scale duplications and repeats.³¹² Indeed, these technologies are enabling the construction of multiple high quality plant genome assemblies^{302, 304, 313-324} and are now poised to increase the number of sequenced plant genomes even further.

Because legumes (family *Fabaceae*) are important in both agriculture and natural ecosystems, primarily due their capacity to form symbiotic relationships with nitrogen fixing bacteria, multiple genome assemblies are now available. Reference assemblies exist for lotus

(*Lotus japonicus*),³²⁵ soybean (*Glycine max*),³²⁶ medicago (*Medicago truncatula*),³²⁷ chickpea (*Cicer arietinum*),³²⁸ mungbean (*Vigna radiata*)³²⁹ and peanut (*Arachis sp.*).^{305, 330} Recently, multiple genome assemblies of a single plant species have begun to appear, enabling the identification of variation in genome content and structure segregating within species,³³¹⁻³³⁵ including legumes.^{331, 334}

Medicago truncatula is a widely studied legume genome, especially in the area of plant-bacterial symbioses. Two *Medicago* accessions have been mainly used for genomic studies, R108 and A17.^{327, 336} The relationship of R108 to A17, the accession used for generating the *M. truncatula* reference genome, makes it valuable both for a technology comparison and as a second *M. truncatula* assembly. Genotype R108 is one of the most distant *M. truncatula* accessions from A17.³³⁷ Relative to A17, R108 has much higher transformation efficiency, has a shorter generation time, and is easier to germinate, making it attractive for genetic studies.³³⁸ Also, R108 is also important to the plant and symbiosis communities because it is the accession that was used to create a large Tnt1-insert population, widely used in functional analysis.^{336, 338} Having two high quality references in *Medicago* therefore allowed us to perform comprehensive genome-scale comparisons between the two assemblies, revealing additional novel R108 sequences as well as increased fine-structure details of important re-arrangement events compared to previous analyses using ALLPATHS-LG assemblies.³³⁴

M. truncatula has a modest genome size, approximately 465 Mb.³³⁹ However, it also has an evolutionary history of whole genome duplications^{340, 341} and frequent local duplications, which appear to be particularly common in this plant species,³²⁷ both of which make assembly difficult. We therefore generated and evaluated five combinations of PacBio, BioNano, and Dovetail technology to see how the technologies could complement each other and to explore

differences in the ordering of technologies. Ultimately, we present a second, high quality reference genome for *M. truncatula* accession R108, based on an optimized combination of the three sequencing/mapping technologies.

Results

Assembly Pb was generated using ~100X PacBio coverage and the FALCON assembler followed by Quiver polishing. Four additional assemblies were then created that had either BioNano (PbBn), Dovetail (PbDt), or both scaffolding technologies added onto the base assembly. The assemblies with both scaffolding technologies were created by applying BioNano and then Dovetail (PbBnDt) or Dovetail and then BioNano (PbDtBn).

Assembly Continuity

The Pb base assembly had just over 1,000 contigs with no gaps in the sequence (Table E.1). It totals just under 400 Mb compared to 412 Mb assembled in the *M. truncatula* A17 reference out of the estimated 465 Mb genome size. The contig N50 for the Pb assembly is 3.77 Mb and the longest sequence is 13.59 Mb. We then added mapping or scaffolding technologies (BioNano and/or Dovetail) on top of this base assembly to improve scaffolding.

Both BioNano and Dovetail (PbBn or PbDt) technologies improved the PacBio only base assembly in similar ways (Table E.1). The number of scaffolds decreased in both assemblies, dropping by 80 scaffolds in the PbBn assembly and 68 scaffolds in the PbDt assembly while having little effect on total scaffold length (Table E.1). The PbBn assembly increased the scaffold length by approximately 1%, adding 4.4 Mb, likely reflecting the fact that BioNano, unlike Dovetail, sizes the gaps it makes when joining sequences. Dovetail adds 100 Ns for each gap it creates, adding only 11.6 kb to the scaffold length.

Table E.1 Number and characteristics of contigs and scaffolds for each of the five assemblies.

	PacBio (Pb)	PacBio BioNano (PbBn)	PacBio Dovetail (PbDt)	PacBio BioNano Dovetail (PbBnDt)	PacBio Dovetail BioNano (PbDtBn)
Assembly software	FALCON	FALCON Irys	FALCON HiRise	FALCON Irys HiRise	FALCON HiRise Irys
Contigs	1,073	1,073	1,121	1,125	1,121
Contig Length	396,973,838	396,973,942	396,973,838	396,973,942	396,973,934
Contig N50 ^a	3,768,504	3,768,512	3,768,504	3,768,512	3,768,504
Scaffolds	1,073	993	1,005	965	942
Scaffold Length	396,973,838	401,421,527	396,985,438	401,429,527	399,955,467
Maximum Scaffold Length	13,488,151	22,885,216	19,275,758	12,137,306	12,557,854
Scaffold N50 ^a	3,768,504	6,819,834	6,895,511	12,137,306	12,557,854

^a N50s were also adjusted to use an assembly length of 400 Mb for all assemblies in order to facilitate comparisons across assemblies. Scaffold and contig N50s adjusted for a 400 Mb assembly size were identical to unadjusted N50s shown above, except for the PbDt scaffold N50 for which the adjusted N50 was 6,348,449 nt.

The scaffold N50s increased substantially for both the PbBn and PbDt assemblies, from 3.8 Mb in the base Pb assembly to over 6.8 Mb in both assemblies (Table E.1). Although the scaffold N50 was slightly higher in the PbDt assembly (6.9 Mb vs 6.8 Mb), the N50 when adjusted for total genome size to allow for comparisons across assemblies (adjusted N50) dropped to 6.3 Mb in the PbDt assembly but remained unchanged in the PbBn assembly. Maximum scaffold sizes increased in both assemblies, from 13.5 Mb in the Pb assembly to 22.1 Mb in the PbBn assembly and 19.3 Mb in the PbDt assembly.

Adding a second technology to the PbBn and PbDt assemblies resulted in two assemblies that differed only in the order in which the BioNano and Dovetail technologies were applied.

Overall, the PbBnDt and PbDtBn assemblies were very similar by scaffold size metrics (Table E.1). Combining all three technologies resulted in slight decreases in the number of scaffolds, slight increases in total scaffold length, and large increases in scaffold N50 (Table E.1). The increase in continuity was particularly striking, with the scaffold N50 nearly doubling to over 12 Mb relative to the PbBn and PbDt assemblies and nearly tripling relative to the Pb base assembly. The maximum scaffold length was slightly larger in the PbBnDt assembly (30.4 Mb vs 27.3 Mb in the PbDtBn assembly), though the PbDtBn assembly had a slightly larger increase over its input assembly (PbDt).

As expected, given that neither BioNano nor Dovetail added a significant amount of sequence data, the number of contigs, contig lengths, and N50s, were nearly identical for all five assemblies (Table E.1). The only substantial change to the contig stats was a slight increase in the number of contigs when Dovetail technology was used, due to the breaking of chimeric contigs (Table E.1).

Assembly Completeness

To assess assembly completeness we examined the number of genomic reads that were captured by the assembly. We used PacBio reads, which were used to create the assemblies, as well as Illumina reads, which represent an independent read set, that were captured by the assemblies. The base (Pb) assembly captured 91.8% of the PacBio reads and 96.8% of the Illumina reads. Moreover, 95.7% of the Illumina reads aligned as pairs with expected orientation and distance, indicating that, at least on the local scale, the assembly is accurate.

Because BioNano and Dovetail are scaffolding technologies, they are not expected to add a substantial amount of additional sequence, but rather to organize the assembly sequences into

longer scaffolds. Indeed, the estimates of assembly completeness obtained through read capture did not change meaningfully upon the addition of these technologies (Supplementary Table S1).

Gene Space Completeness

In order to investigate the completeness of the gene space in the five assemblies we determined rates of capture for conserved single-copy eukaryotic genes (BUSCO)³¹ and an R108 transcriptome assembly, and assessed MAKER-P annotations. Because completeness results for all 5 assemblies were quite similar, we discuss only results for the Pb base assembly and present results for the other assemblies in the supplement (Supplementary Table S2). The BUSCO analysis indicates that the base assembly (Pb) captured nearly all of the genes (878 of the 956 genes in the dataset; 91.8%). Nearly 16% (151) of the putative single-copy genes in the BUSCO database were duplicated within the assemblies. These putative duplicates might be due to true duplications in the R108 genome or they might be due to artificial redundancy in the assembly. Even though the BUSCO gene groups are generally single copy, given plant genome duplication rates it isn't surprising that some of the genes are duplicated.

In addition to looking at capture of conserved genes, we also looked at capture of an R108 RNA-Seq assembly that was produced independently of the genome. Assembly completeness results were similar to those seen with BUSCO, with approximately 92% (94,519) transcripts captured. However, as would be expected, the duplication rate was much higher than that seen in BUSCO, which specifically focuses on single copy genes. In the R108 transcript assembly, 37,929 transcripts (37% of total, 40.1% of aligned transcripts) were duplicated.

Finally, we analyzed the total number of genes predicted from MAKER-P. There were 54,111 genes compared to 50,894 gene loci in Mt4.0 (accession A17). This gives additional

confirmation that the gene space is largely complete. Further, there may be additional genes in the R108 Pb assembly not found in A17 (see below).

Joins and Breaks

When characterizing the joins made by BioNano and Dovetail, some interesting trends emerged (Supplementary Table S3). Dovetail joined more scaffolds when applied to the base (Pb) assembly compared to BioNano. Dovetail joined 172 Pb scaffolds into 64 PbDt scaffolds while BioNano joined 140 Pb scaffolds into 50 PbBn scaffolds. The same trend of more joins for Dovetail compared to BioNano held when adding a second scaffolding or mapping technology. Dovetail joined 114 PbBn scaffolds into 45 PbBnDt scaffolds and BioNano joined 96 PbDt scaffolds into 33 PbDtBn scaffolds. For the two contrasting assemblies created with all technologies, the two rounds of scaffolding resulted in a total of 254 scaffolds joined in the PbBnDt assembly and 268 scaffolds joined in the PbDtBn assembly, a difference of just over 5%. While Dovetail joined more scaffolds, BioNano had a higher average number of scaffolds per join (Supplementary Table S3).

To determine the characteristics of scaffolds that were being joined, we pulled out scaffolds from the input assembly that were joined by either technology in either round (Table E.2, Supplementary Table S4). The biggest difference between the two technologies was in the ability to join shorter scaffolds. Dovetail was able to join scaffolds as short as 4,765 nucleotides into a larger super-scaffold (in both rounds 1 and 2), whereas the minimum scaffold size that BioNano was able to join was 172,295 in round 1 and 98,093 in round 2. To further understand the ability of Dovetail to join smaller contigs, we quantified the number of input scaffolds less than 100kb that each technology was able to join (Supplementary Table S4). Dovetail joined 35 sub-100kb scaffolds (17 in round 1 and 18 in round 2). BioNano, on the other hand joined only 1

sub-100kb scaffold total (in round 2), and that scaffold was nearly 100kb (98,093 nt). Clearly, Dovetail is better at incorporating short scaffolds less than 100 kb.

While Dovetail appears to be better at incorporating shorter scaffolds, it also appears to more effectively join longer scaffolds. When only scaffolds ≥ 100 kb cutoff were examined, Dovetail joined 253 input scaffolds and BioNano joined 237 across both rounds. Similarly, when only very large scaffolds were examined (≥ 1 Mb) Dovetail joined 141 input scaffolds and BioNano joined 128 across both rounds. Dovetail had a higher number of joins at each cutoff when the data were broken down by each round as well (data not shown).

Table E.2 Characteristics of Input Scaffolds that were Joined by BioNano and/or Dovetail.

Assembly	Pb -> PbDt	Pb -> PbBn	PbDt -> PbDtBn	PbBn -> PbBnDt
Scaffolds	172	140	96	114
Max Scaffold	13,488,151	13,488,151	19,275,758	22,885,216
Scaffold N50	3,957,684	3,698,567	6,895,511	6,819,834
Scaffold N90	854,372	929,179	1,425,957	1,427,073
Min Scaffold	4,765	172,295	98,093	4,765
Total Scaffold Length	307,402,024	293,002,927	260,974,793	289,680,947

To identify similarities between the two technologies, we determined whether some of the joins made were the same between BioNano and Dovetail. We focused on the first round, where each technology was added onto the Pb assembly, looking for cases where the same Pb scaffolds were joined into a super-scaffold. There were 47 Pb input scaffolds that were scaffolded by both BioNano and Dovetail, resulting in 21 scaffolds in the PbDt assembly and 20 scaffolds in the PbBn assembly. The fact that these joins were made by two independent technologies improves our confidence in these joins. Given that there were also joins made that were unique to both technologies supports the increased continuity and additional joins that we are seeing in assemblies that have both technologies added.

In order to determine whether Dovetail was breaking apart scaffolds that BioNano had previously created by merging Pb scaffolds, we looked further into the Dovetail breaks. In other words, we asked whether any of the joins made by BioNano when generating the PbBn assembly were subsequently split by Dovetail when applied to the PbBn assembly to generate the PbBnDt assembly. From the merged scaffolds generated in the PbBn assembly, only 8 PbBn scaffolds were broken by Dovetail in the PbBnDt assembly and no breaks occurred directly inside the gaps that had been generated by BioNano (median distance from gap was 137,686 nt). We generally found read support spanning these regions, with half or more of the alignments having equally good hits to other regions of the assembly (data not shown). This indicates that these were large repetitive regions and it was difficult to say confidently whether the region should be joined (BioNano correct) or broken (Dovetail correct).

Joins and Breaks in Relation to A17

We used alignments of first round assembly scaffolds (PbBn and PbDt) to A17 to predict whether scaffold joins were correct. If joined pieces of a scaffolds mapped to the same A17 chromosome, this lends support for the join. Because of the evolutionary distance between R108 and A17, rearrangements are expected, so a negative result doesn't necessarily mean the join is incorrect. However, vastly different rates of A17 synteny between scaffold joins made by BioNano and Dovetail would suggest better accuracy for one of the technologies.

Scaffolds joined by BioNano mapped to the same A17 chromosome at a rate of 78.57% while those joined by Dovetail mapped to the same A17 chromosome at a rate of 93.75%. This suggests that Dovetail had a better accuracy than BioNano. Scaffolds with joins that were supported by both BioNano and Dovetail appear to be of higher accuracy based on alignments to A17. For BioNano, while over half of joins (54.54%) were from scaffolds that had similar joins

by Dovetail, only 20.00% of joins that mapped to different A17 chromosomes were supported by a similar Dovetail scaffold. This resulted in a 90.91% of Dovetail-supported BioNano joins that mapped to the same A17 chromosome, an increase of 12.34% over all BioNano joins. Dovetail, had more joins than BioNano (see above), with 36.67% of the joins supported by a similar BioNano scaffold. A similar percentage was seen in the number of BioNano-supported Dovetail joins compared to all Dovetail joins (33.33%), resulting in 94.29% of BioNano-supported Dovetail joins aligning to a single A17 chromosome, representing an increase of 0.54%.

Finally, we looked at A17 synteny in the eight PbBn scaffolds that were subsequently broken by Dovetail in the PbBnDt assembly. Three of the scaffolds had input pieces that mapped to chromosome U (unknown), making it difficult to determine A17 synteny and indicating that repetitive sequence is likely that made it difficult to make a chromosome assignment. Of the other 5 scaffolds, 3 mapped to the same A17 chromosome, supporting the BioNano join and 2 mapped to different chromosomes, supporting the subsequent Dovetail break.

Gaps

The sizing of gaps in BioNano versus the addition of 100 nts in Dovetail, resulted in an increase in the amount of nucleotides added to the total scaffold length in the first round for BioNano compared to Dovetail (Table E.1).

Table E.3 Characteristics of the gaps introduced into the assemblies by BioNano and Dovetail. Note, there are no gaps in the Pb only base assembly so it is not included.

	PbBn	PbDt	PbBnDt	PbDtBn
Captured Gaps	80	116	160	179
Max Gap	647,836	100	647,836	647,022
Min Gap	500	100	100	100
Mean Gap	55,595	100	27,847	16,657
Gap N50	171,515	100	171,515	105,896
Total Gap Length	4,447,585	11,600	4,455,585	2,981,533

In order to see how the gap strategies of BioNano and Dovetail interact, we analyzed the second round assemblies (PbBnDt and PbDtBn), which have both technologies incorporated but with differing order. When a second scaffolding or mapping technology was added to an assembly that already incorporated the other technology, the gaps from the first technology were carried over intact. As noted above, Dovetail sometimes broke apart scaffolds that BioNano had put together. However, when breaking these scaffolds, Dovetail never broke the scaffolds within the gap generated by BioNano but rather broke it in a nearby position. In assemblies where BioNano was added to the PbDt assembly, the minimum gap size that BioNano introduced was 500 nt. This minimum size might be because 500nt is the minimum gap BioNano can span. Alternatively, given that the assemblies are all based upon PacBio data, it may be that smaller gaps were easily bridged by the PacBio data itself.

The assemblies with both BioNano and Dovetail (PbBnDt and PbDtBn) ended up with a similar number of captured gaps (Table E.3). The maximum gap length was over 647 kb, generated when adding BioNano onto the Pb assembly. Although Dovetail doesn't size its gaps, given the insert size of ~100kb, it is likely that most of the gaps fall below this range. BioNano, with a gap N50 of 171,515 (Table E.3), therefore was able to jump across larger distances than Dovetail.

A similarly sized gap generated when adding BioNano onto the PbDt assembly traces back to the same Pb scaffolds as the join made by BioNano on the Pb assembly. Finally, the total gap length varies. Among those assemblies that contain sized gaps (PbBn, PbBnDt, and PbDtBn), the PbDtBn assembly has considerably fewer nts in gaps compared to the other two. This is somewhat surprising given the fact that this assembly has the most gaps of any assembly and that there were more joins made over the two rounds in the PbDtBn assembly (268) than

over both rounds in the PbBnDt assembly (254) (Supplementary Table S3). Overall, the gap sizes in PbDtBn are smaller (Table E.3), accounting for the lower number of nts in gaps.

Finally, in order to surmise the nature of sequence in the gaps and why contigs stop instead of continuing on, we looked at the sequence flanking the gaps (10kb). Interestingly, the joins made by BioNano and Dovetail (and the breaks made by Dovetail) were enriched for repetitive sequence in the regions flanking the gap introduced with the join (Supplementary Figure S1). BioNano and Dovetail both appear to be able to jump across larger repetitive regions than is possible with PacBio reads. In other words, the value of the two technologies is often in their ability to bridge across repetitive regions that PacBio reads cannot currently cross.

Ordering of Technologies

The ordering of the scaffolding or mapping technologies made a difference to the continuity and completeness statistics (Table E.1, Supplementary Tables S1 and S2). Using Dovetail before BioNano provides multiple benefits. The fact that Dovetail breaks chimeric scaffolds automatically means that using it up front provides a cleaner assembly template for BioNano. Dovetail's ability to scaffold much smaller pieces of DNA compared to BioNano means that if Dovetail is used up front, more joins will be made and a better base sequence assembly constructed.

Final Assembly Draft

In order to create the best reference assembly, we gap-filled the PbDtBn assembly using PBJelly (named R108 version 1.0, Table E.4). The PbDtBn assembly was chosen because it had slightly better assembly stats compared to PbBnDt (Table E.1, Supplementary Tables S1 and S2). For the five preliminary assemblies interrogated above, we did not do any gap filling or polishing (except that the base assembly was polished with Quiver) because these methods

would obscure the effects that the BioNano and Dovetail technologies were having on the assembly process. Nevertheless, PBJelly was used for gap-filling as well as super-scaffolding on the final assembly draft in order to improve continuity. While gap filling can be over-aggressive especially if flanking sequences are repetitive, having some sequence, even if not perfect, is often better than having just Ns. In addition, using Dovetail and then BioNano enabled us to use independent data to bring scaffolds together and size the gap between them, making us more confident with doing gap-filling.

Table E.4 Assembly Statistics for R108 version 1.0 (PbDtBn PBJelly gap filled) and its input assembly (PbDtBn).

	R108 v 1.0	PbDtBn
Contigs	1,016	1,121
Contig Length	399,348,944	396,973,934
Contig N50	5,925,378	3,768,504
Scaffolds	909	942
Scaffold Length	402,065,285	399,955,467
Scaffold N50	12,848,239	12,557,854

PBJelly was able to fill many of the captured gaps, increasing the continuity of the PbDtBn assembly (Tables E.1 and E.4). In total, it filled in 415 of 522 gaps (79.50%). As expected, gap-filling was able to fill far more small than large gaps, resulting in an increase of the gap N50 from 12,335nt to 110,194nt, a nearly 9-fold increase. The latter is much longer than typical PacBio reads and may represent repeats that were too long to span with these reads. The total gap length was only reduced by 8.82% despite the fact that 79.50% of the gaps were filled, again reflecting the preferential filling of small gaps. Nevertheless, continuity is much improved. The number of contigs dropped by ~ 12% to just over 1000 (1016 contigs), and the contig N50 increased from 3,768,504nt to 5,925,378nt, representing an increase of 57.23%. Gap filling had little effect on the number of scaffolds, scaffold N50, or total assembly size (differences between gap filled and ungapped assemblies were < 0.5%).

The completeness stats of the gap filled assembly improved slightly relative to the PbDtBn assembly before gap-filling (Supplementary Tables S1 and S2). The final draft R108 v 1.0, assembly captured 93.2% of Pb reads and 96.8% of Illumina reads. Of the original Illumina readset, 95.8% were not only mapped but also properly paired, indicating that the assembly has captured most of the genome. The R108 v 1 assembly has captured most of the gene space, with estimates ranging from 92.3% for the transcript assembly to 95.2% for the BUSCO assembly, and 55,706 genes predicted MAKER-P. Overall, this final draft of the R108 assembly captures nearly all the assembly and gene space.

Novel sequences revealed by the R108 assembly

A new high quality reference sequence for R108 allowed a side-by-side comparison of two *Medicago* accessions (A17 and R108). We were able to build chromosome-level synteny blocks between R108 and A17. We also found extensive novel sequence in the R108 assembly that was not part of the A17 reference assembly (Table E.5). There was nearly 23 Mb of R108 assembly sequence that could not be found in the A17 assembly. This represents 5.7% of the nucleotides in the R108 genome. These “novel” sequences are likely a mix of sequences that are truly novel in the R108 genome as well as sequences that are present in both genomes but have diverged beyond our ability to detect them or sequences that are in the A17 genome but didn’t make it into the A17 assembly. Out the nearly 23 Mb of novel R108 sequence, 1.6 Mb represent novel R108 coding sequence that could not be found in the A17 assembly, values quite similar to those observed with an earlier ALLPATHS-LG³⁴² assembly of R108³³⁴. These regions contain candidate R108-specific genes or gene that were deleted from A17 or arose independently in the R108 lineage. Table E.5 R108 v 1.0 assembly characteristics in comparison to the A17 reference assembly.

Table E.5 R108 v 1.0 assembly characteristics in comparison to the A17 reference assembly.

	Nucleotides	% Nucleotides
Total Bases	399,348,955	100.00%
Repetitive	96,760,262	24.23%
Alignable to A17	366,489,898	91.77%
Bases in Synteny with A17	283,853,354	71.08%
Novel Sequences vs A17	22,763,508	5.70%
Novel Coding Sequences vs A17	1,623,097	0.41%

Chromosomal-scale translocation

Although R108 is phylogenetically distant from A17 compared to other accessions, we were able to align more than 280 Mb of syntenic regions in both genomes (Table E.5), representing over 70% of the R108 assembly. These numbers also correspond well with sequence comparisons based on an earlier ALLPATHS-LG assembly of R108.³³⁴ Within these synteny blocks, extensive variations were discovered including single nucleotide changes, small insertions and deletions, as well as large structural changes such as inversion and translocation. While most structural changes were TE-related and only involve small local regions, we identified two large rearrangements on chromosomes 4 and 8 between R108 and A17. Through synteny comparison, we found one R108 scaffold (scf005, 16.4Mb) spanning the upper arm of chromosome 4 and the lower arm of chromosome 8 in A17, and another two scaffolds (scf015, 12.0Mb and scf002, 17.6Mb) together spanning the upper arm of chromosome 8 plus the lower arm of chromosome 4 (Figure E.1), indicating a chromosomal-scale translocation between the reference *Medicago* accession (A17) and the widely-used R108 accession.

Previously, Kamphuis *et al.* reported a rearrangement between linkage groups 4 and 8 in the reference accession A17 relative to other accessions.³⁴³ Using genetic markers and linkage mapping, the authors hypothesized a chromosomal-scale translocation private to A17 which

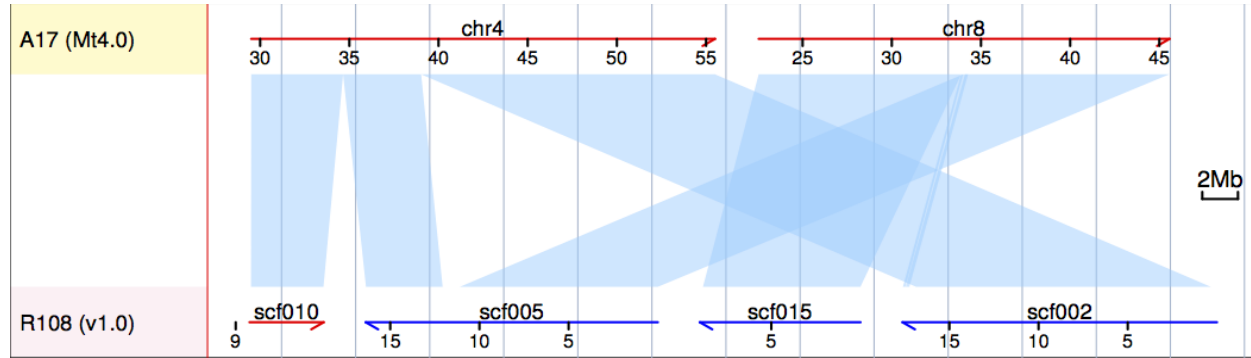


Figure E.1 Synteny alignment of partial chromosomes 4 and 8 between A17 and R108 confirms rearrangement of the long arms of the chromosomes.

involves the lower arms of chromosomes 4 and 8.³⁴³ To date, however, the physical location of the rearrangement has not been determined and, in fact, the rearrangement itself has not been elaborated through genome sequencing. Lack of high quality genome assemblies of non-A17 accessions certainly hindered such whole genome comparison. However, even with the whole genome assemblies available (including the earlier R108 ALLPATHS-LG assembly), it is still difficult to fully resolve rearrangement events at such chromosomal scale given the relatively short scaffold span of most sequencing and assembly techniques. Figure E.2 clearly illustrates the improvements in resolving large-scale structural variation using long PacBio reads together with scaffolding or mapping technologies such as Dovetail and BioNano, over traditional Illumina-based assembly or assembly based on PacBio reads alone. Using the same synteny pipeline we aligned the Illumina-based R108 assembly, assembled with ALLPATHS-LG,³⁴² to A17. The rearrangement region (~50Mb) on chromosomes 4 and 8 was split into ~30 independent scaffolds in the ALLPATHS-LG R108 assembly (Figure E.2, top panel). The PacBio-based assembly (Pb), on the other hand, captured the region in ~10 scaffolds and partially resolved the breakpoint on chromosome 4 (Figure E.2, middle panel). With the aid of BioNano and Dovetail technologies, the affected region was captured in four long scaffolds in the final R108 assembly (PacBio+Dovetail+BioNano) with all breakpoints clearly resolved

(Figure E.2, bottom panel). We were able to pinpoint exact breakpoints of the translocation to a single region on chromosome 4 and three regions on chromosome 8, something that could not be done with the Illumina-based ALLPATHS-LG assembly (Figure E.3). Interestingly, each of the four breakpoints involves a gap (i.e., 'N's) in the A17 reference, with one 7.5 kbp gap and three 100 bp gaps, the latter representing gaps of undetermined size (Haibao Tang, personal communication). These gaps indicate that the regions in and around the rearrangement breakpoints are structurally unstable, repetitive and/or difficult to assemble even using a BAC-by-BAC approach. We found numerous transposable element genes near the breakpoints, including a reverse transcriptase, a GAG-pre integrase and a cluster of 6 transferases near breakpoint 1, two helicases around breakpoint 2, two retrotransposons (UBN2) and two reverse transcriptases around breakpoint 3, and a MULE transposase right next to breakpoint 4. Intriguingly, a cluster of at least 10 CC-NBS-LRRs was found both upstream and downstream of breakpoint 2, and two CC-NBS-LRRs were also found right next to breakpoint 3, possibly suggesting a structural role of these resistance genes in plant genomes.

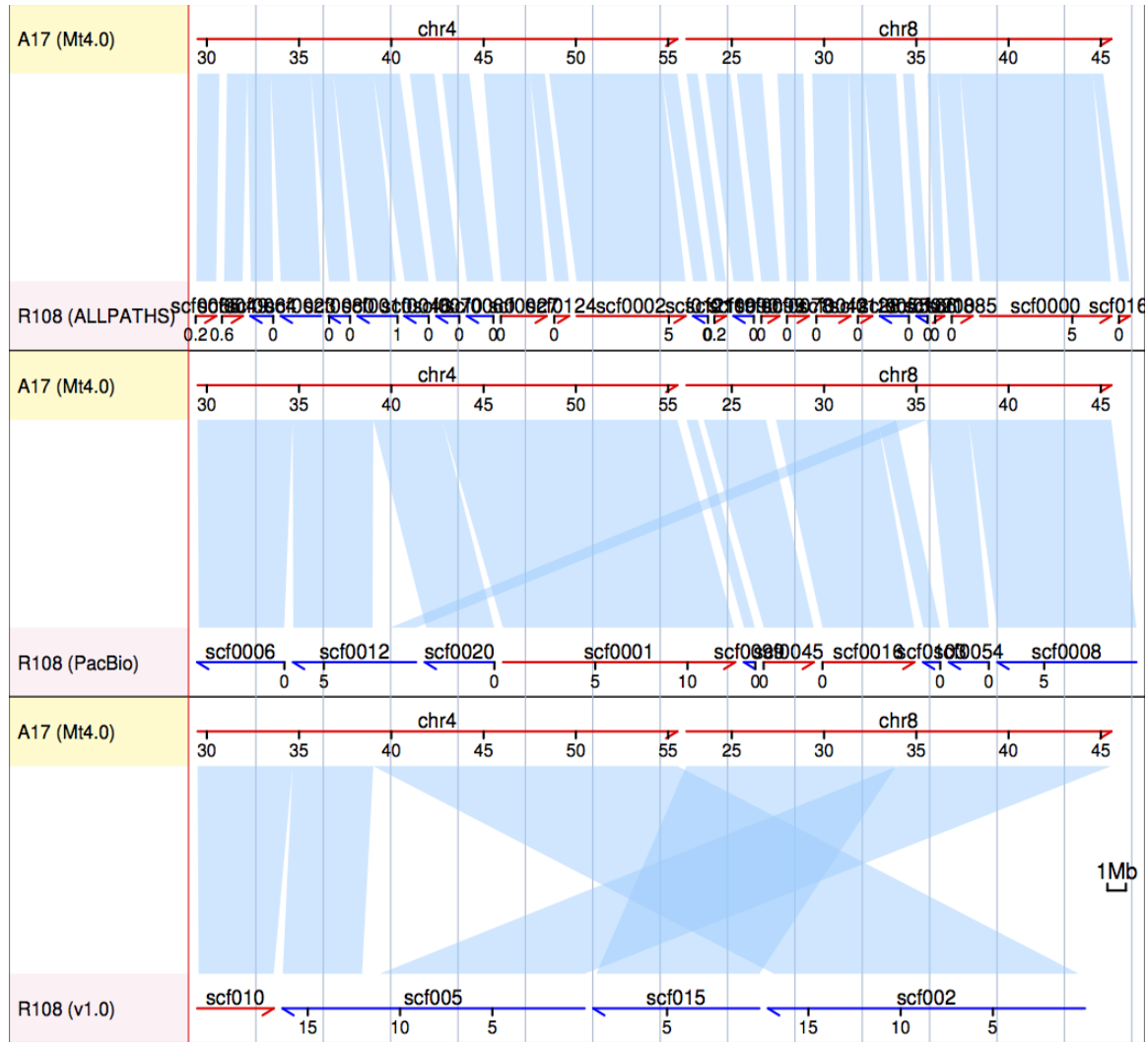


Figure E.2 Synteny alignment of partial A17 chromosomes 4 and 8 against syntenic regions in the R108 Illumina-based assembly (top panel), PacBio-based assembly (Pb, middle panel) as well as the gap-filled PbDtBn (v1.0) assembly (bottom panel).

In addition to the translocation, we noticed two large stretches of R108 sequences (1.15 Mb and 430 Kb) downstream from the translocation breakpoints on chromosome 4 and 8 (Figure E.3 red segments) that didn't have a syntenic match in A17. The chromosome 4 insertion in R108 is a ~1 Mb region with no synteny to A17 and right next to the chr4-8 translocation breakpoint. Both the translocation and insertion are found in several other accessions including HM034 and HM185 using a similar synteny comparison approach (data not shown). It is thus

likely that the translocation is private to A17, which is consistent with,³⁴³ and this large insertion in R108 actually represents a private deletion in A17 which is expected to be found in the majority of *M. truncatula* accessions. Further examination revealed that most of the insertion is novel. A total of 623 kbp of novel segments that do not align anywhere in A17 were identified in this region with 136 genes found in this region (Supplementary Table S5).

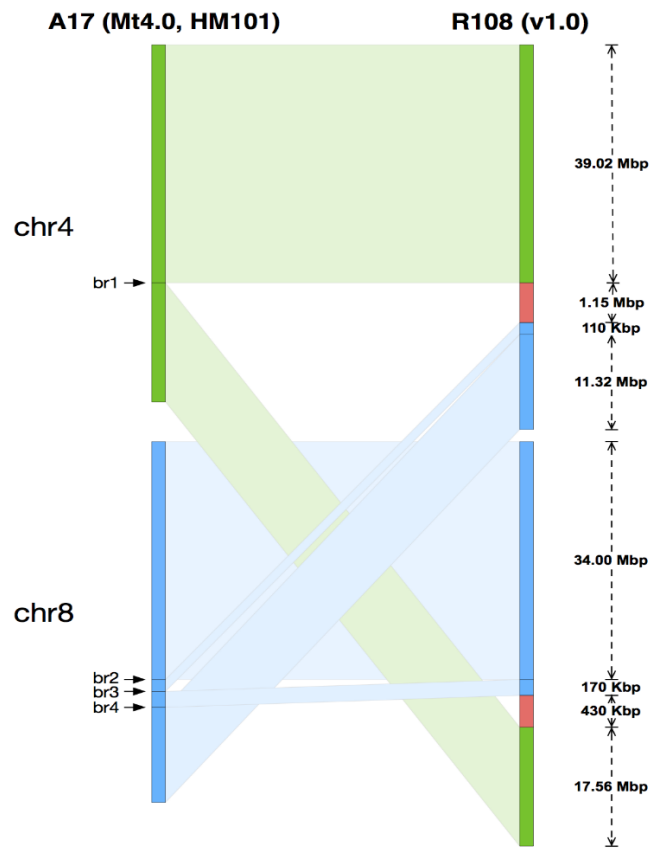


Figure E.3 Schematic of the rearrangement between chromosomes 4 and 8 in A17 (left) compared to R108 (right). Green segments indicate homology to A17's chromosome 4 while blue segments indicate homology to A17 chromosome 8. Red segments indicate sequences not present in the A17 reference). Breakpoint 1 (br1) is pinpointed to a 104 bp region (chr4:39,021,788-39,021,891) and includes a 100 bp gap. Breakpoint 2 (br2) is pinpointed to a 7,665 bp region (chr8:33,996,308-34,003,972) and includes a 7,663 bp gap. Breakpoint 3 (br3) is pinpointed to a 708 bp region (chr8: 34,107,285-34,107,992) and includes a 100 bp gap. Breakpoint 4 is pinpointed to a 277 bp region (chr8:34,275,249-34,275,525) and includes a 100 bp gap).

Discussion

This work represents the first published example we are aware of examining multiple next generation scaffolding and mapping technologies in all possible combinations with a comparative analysis of their contributions. PacBio long reads combined with BioNano and Dovetail technologies have allowed us to generate a second, reference quality assembly for the model legume, *M. truncatula*, in the functionally-important R108 accession. In the process, we discovered important insights into how these technologies overlap and complement each other enabling us to propose an optimal strategy for their incorporation.

Novel Sequence Was Found in the R108 Assembly

Long reads improve the continuity of assemblies.^{315, 344-348} However, continuity is only one advantage of using long reads. The long reads help to correctly capture ambiguous regions of the genome in the assembly, including repeats and tandemly duplicated genes. Locally duplicated genes can be especially problematic as they are often collapsed or over-expanded in Illumina-only or even Illumina/PacBio hybrid assemblies (Miller et al, submitted). Using PacBio long reads, therefore, results in capture of additional sequence that is not possible with short reads. In addition, we capture accession specific sequences as well. In total, over 22 Mb of novel sequence, including 1.6 Mb of coding sequence were identified.

Technologies Made Similar Continuity Gains and Are Valuable Individually

Similar continuity gains were made by each technology in each round, as was seen in.³¹³ Both technologies improved the base Pb assembly, improving the 3.8 Mb scaffold N50 of the Pb assembly to just over 6.8 Mb (Table E.1). Indeed, many of the same joins were made between both of the technologies. Both technologies, individually, were valuable in increasing continuity.

Despite the challenges of assembly the *M. truncatula* genome, with its history of whole genome duplication and high rate of locate duplication, there are many plant genomes that are much more complicated than the 500 Mb, largely homozygous *Medicago truncatula* genome. Increases in genome size, repetitive content, and the number of tandem, segmental, or whole genome duplications will change the dynamics of the assembly and the contributions of the technologies. In *Medicago* described here, the PacBio assembly came together quite well, making the improvements when using BioNano and Dovetail less dramatic than they might have been. As genome complexity increases, including repeat and duplication content, coherent PacBio assemblies become increasingly difficult. As PacBio assemblies become more fragmented with increased genome complexity, we expect that the improvement in the assembly when adding BioNano and/or Dovetail will become increasingly crucial, leading to greater relative improvements, even while becoming more challenging. The assembly improvement with both technologies should follow similar patterns with increased genome complexity until extremely high levels of complexity, especially repeat size, become limiting even for these technologies.

Further Gains Were Made Using Both Technologies

Though similar gains were seen when using either scaffolding or mapping technology, the use of both technologies together increased continuity gains and join numbers further (Table 1 and Supplementary Table S3).³¹³ With a combined approach the two technologies were complementary by enabling additional joins than either Dovetail or BioNano could make independently. Using both scaffolding technologies in either order (PbDtBn or PbBnDt) increased the scaffold N50 to just over 12.1 Mb (Table E.1).

One explanation for the complementarity between the two technologies may be a function of the differences in biases of the two technologies. BioNano's information content is in restriction sites and the distances between them. As such, BioNano is highly dependent on the motif density of the restriction enzymes used,^{349, 350} which can vary within a genome (Supplementary Figure S2A). Genomic regions where motif density is high become “fragile sites”, that destabilize the DNA and resulting in limited or no coverage in the maps, and breaks in the genome map contigs.^{26, 303, 310, 350} In these regions scaffolding of the assembly simply cannot occur. By contrast, regions of the genome with too low of a density of cutting sites also will result in low label density and missed join opportunities (a minimum of eight restriction sites is required in each DNA molecule, which is a minimum of 150 kb).

Dovetail is based on Hi-C technology, an extension of chromosome conformation capture, which has its own documented biases.^{351, 352} Dovetail's information content is “contact probabilities,” indicating the probability that any two regions in the genome will be brought together during the ligation stage and is inversely correlated with distance. Dovetail, which incorporates Illumina sequencing, also inherits biases in next generation sequencing and alignment, such as biases in the amplification, shearing and mapping steps.

Join Accuracy Appears to be Higher in Dovetail Compared To BioNano

Using A17 synteny as a proxy for accuracy of joined R108 scaffolds, Dovetail had a much higher percentage of joins mapping to the same A17 chromosome compared to BioNano (93.75% vs 78.57%), suggesting that accuracy is higher in Dovetail than in BioNano. Further, when looking at joins in scaffolds supported by both technologies, Dovetail-supported BioNano joins mapped to the same A17 chromosome 90.91%, an increase of 12.34% over all BioNano joins. This suggests that Dovetail confirmation increases the accuracy of BioNano joins.

BioNano-supported Dovetail joins, however, increased mapping to the same A17 chromosome by only 0.54%, suggesting that BioNano confirmation did little to improve Dovetail accuracy.

These data argue that Dovetail joins are more accurate than BioNano joins. However, we cannot rule out that the possibility that the larger distances that the BioNano technology spanned while joining scaffolds (described above) might make it less likely that two joined scaffolds fall into a region that is syntenic with A17 given that synteny tends to decrease with distance. BioNano-joined scaffolds, therefore, might map to multiple A17 chromosomes more than Dovetail-joined scaffolds due to synteny breakdown rather than inaccuracy of joins. However, given that BioNano gaps span less than 200kb and that the majority of the R108 genome has synteny blocks with A17 that are greater than 1 Mb (Figures 1-3),³³⁴ we expect this difference to be small and the difference between Dovetail and BioNano join accuracy to be real.

Alternatively, Dovetail breaks performed much worse than joins using A17 synteny as a measure. Of the PbBn scaffolds subsequently broken by Dovetail in the PbBnDt assembly, only 40% of them mapped to different A17 chromosomes, indicating that Dovetail might be breaking more correct BioNano joins than incorrect ones.

A17 chromosomal mapping is far from a perfect gold standard given the evolutionary distance between A17 and R108. Joined segments of R108 scaffolds that map to different A17 chromosomes may still map to the same R108 chromosome. Indeed, one of the joins shared by both Dovetail and BioNano that mapped to different A17 chromosomes corresponds to the known chromosome 4/8 translocation. This join, therefore, is correct, even though synteny to A17 put it on two different chromosomes. It is possible that there are other regions where synteny to A17 doesn't accurately predict synteny in R108. Using long-range physical information, such as Hi-C data or a genetic map that involves R108, could allow us to better

validate the BioNano and Dovetail technologies as well as to obtain chromosome-scale ordering of the genome assembly.

Strengths and Weaknesses Dictate Strategy for Ordering Technologies

For the final assembly, we chose to gap-fill the PbDtBn assembly rather than the PbBnDt assembly. This decision was based not only on comparisons of important assembly continuity and completeness statistics, as described above, but also on the knowledge we uncovered about the differences between the scaffolding and mapping technologies.

One important difference between the two technologies is their ability to incorporate smaller scaffolds. In our study, Dovetail incorporated thirty-five small scaffolds (less than 100 kb) over both rounds but BioNano incorporated only one. The minimum scaffold size joined by BioNano (98.1 kb) was more than 20 times larger than the minimum scaffold size joined by Dovetail (4.8 kb). Similar results were found when applying BioNano maps to the short arm of wheat chromosome 7D where the optimum size for incorporation by BioNano was 90 kb or higher³⁵⁰ and sequences shorter than 30 kb could not anchored reliably. Given that the scaffold N50 was 3.7 Mb in the Pb assembly to which these technologies was added, the discrepancy between the two technologies in joining scaffolds less than 100 kb did not have as great an effect on our assemblies. However, if a much more fragmented assembly were used, we would expect Dovetail to perform much better than BioNano if only one scaffolding or mapping technology were used. If both technologies are used, applying Dovetail first to incorporate the smaller scaffolds and create a more contiguous substrate for BioNano to use makes sense and would be especially critical for highly fragmented assemblies.

A second difference in the two technologies also supports applying Dovetail prior to BioNano for combined strategies. Dovetail breaks sequences it identifies as chimeric as it runs

the software. BioNano logs potential chimeric sequences, but does not induce breaks in the assembly without manual intervention. Hence, if BioNano is applied first, chimeric contigs may not yet be properly separated when the assembler's master plan for scaffolding is being formed. Having a more accurate assembly up-front, as should occur when Dovetail is applied first, is always best before scaffolding assemblies.

Both technologies were able to bridge larger duplicated and/or repetitive regions than was PacBio, which requires multiple reads long enough to span an ambiguous region. With only ten percent of the sequenced nts in PacBio reads longer than 18,555 nt (N10), the ability of PacBio to span ambiguous regions is likely limited to a similar size, though longer reads will increase the size of the spannable repeats. Therefore, both mapping technologies can add value for spanning ambiguous regions that are beyond the reach of current PacBio capabilities. However, both technologies are limited in the size of gap they can span. Dovetail is limited by its longest pairs, which in this study, likely kept joins to around 100kb or less, though without sized gaps it is difficult to figure out the true maximum. BioNano can join scaffolds over much larger gaps. The largest span made in this study created a gap of nearly 650kb, though most joins spanned less than 100 kb (Table E.3). Nevertheless, Dovetail and BioNano both were able to span ambiguous regions that were beyond PacBio's current capability.

Conclusions

The use and analysis of both BioNano and Dovetail technologies in all possible combinations is novel and yielded strategic information about how best to apply these strategies to PacBio. Both technologies were able to span repetitive regions that PacBio was unable to bridge. Using PacBio, followed by Dovetail and then BioNano, and then gap-filled with PBJelly, we have generated a second, reference quality assembly for *M. truncatula*. Because of the

distance between R108 and the A17 reference as well as the inability to interbreed them to create a genetic map, having a second high quality *M. truncatula* reference has been a priority in the *Medicago truncatula* community. A second reference assembly has yielded novel sequence and will be an important resource for the R108 functional community to support gene-finding in the *Tnt1* lines. The R108 reference assembly has also allowed us to investigate the details of the A17 translocation.

Methods

We generated five genome assemblies: a PacBio only assembly (Pb), a PacBio base assembly that was scaffold together with either Dovetail (PbDt) or BioNano (PtBn), a Pb base assembly that was scaffold together with Dovetail and then BioNano (PbDtBn) and a Pb base assembly that was scaffold together with BioNano and then Dovetail (PbBnDt). The completeness of each assembly was evaluated by alignments of PacBio reads as well as independent Illumina reads, and capture of an independent transcriptome as well as core eukaryotic genes. For comparison, we used the A17 version 4.0 reference genome.³³⁹

PacBio Sequencing and Assembly

DNA for PacBio assemblies was obtained from fifty grams of young leaf tissue obtained from multiple plants grown in the greenhouse and dark-treated for 24 hours. High molecular weight genomic DNA was generated by Amplicon Express (Pullman, WA) using their standard BAC nuclei prep followed by a CTAB liquid DNA precipitation.

Whole-genome DNA sequencing was performed using a Pacific Biosciences RS II instrument (Pacific BioSciences, Menlo Park, CA). Libraries were constructed using the PacBio 20-Kb protocol.¹⁷² These libraries were loaded onto 122 SMRT cells and sequenced using P4/P6

polymerase and C2/C4 chemistry with 3- and 6-hour movie times, respectively. PacBio sequencing yielded approximately 107X sequence coverage. A *de novo* assembly of PacBio reads was generated using FALCON³¹⁵ assembler version 0.4 using default parameters. Contigs smaller than 1kb were removed. In order to improve the accuracy of the assembly, Quiver polishing was done on SMRT portal (version smrtanalysis_2.3.0.140936.p5.167094) using the "RS_Resequencing" protocol using the latest version available at the time.

Dovetail

DNA from Amplicon Express (described above) was used. A Chicago library (Dovetail Genomics LLC, Santa Cruz, CA)²⁹ was generated using the *DpnII* restriction endonuclease (GATC). Briefly, this entailed reconstituting chromatin using purified histones and chromatin assembly factors, followed by cross-linking the chromatin using formaldehyde. DNA was then digested using the *DpnII* restriction endonuclease. The resulting sticky ends were filled in with thiolated and biotinylated nucleotides. A blunt end ligation of free ends followed by removal of the crosslinking and proteins yielded fragments with DNA joined across distances of up to about 100 kb. An exonuclease was used to remove the biotinylated nucleotides. The thiolated nucleotides, which were proximal to the biotinylated nucleotides, protected the DNA from further exonucleation.

The resulting DNA fragments were taken through a standard Illumina library prep, including shearing and adapter ligation. The library was sequenced on an Illumina HiSeq 2000 (2x100 Base Pairs) to a physical coverage level of ~588X (67X sequence coverage).

Sequence data generated from this library were used to scaffold the PacBio *de novo* assembly through Dovetail's HiRise™ pipeline v. 1.3.0-57-g4d1fc9b.²⁹ In short, Chicago library reads were mapped back to the assembly using a modified version of SNAP

(<http://snap.cs.berkeley.edu/>). Pairs in which both reads were uniquely mapped were used to generate a likelihood model representing how chromatin crosslinking brings sequences together. A graph where the nodes are contigs and the edges are ordered integer pairs representing placement of the paired reads in the contigs was used for scaffolding beginning with high confidence linear subpaths and prioritizing joins in order of log likelihood improvement. During the process, in addition to joining sequences, putative chimeric sequences were broken. An iterative approach was taken by feeding the resulting scaffolds back into the pipeline. Refinement of local ordering and orientation and gap closing using Meraculous's Marauder module was done at the end [60].

BioNano

Five grams of young leaf tissue was obtained from greenhouse-grown plants dark-treated for 24 hours before harvest. High molecular weight DNA was extracted and a *de novo* whole genome map assembly was generated using the BioNano Genomics (BNG) (BioNano Genomics, San Diego, CA) platform at the Bioinformatics Center at Kansas State University. High Molecular Weight (HMW) DNA was nicked and labeled according to the IrysPrep protocol. In brief, HMW DNA was double digested by a cocktail of single-stranded nicking endonucleases, *Nt.BspQI* (GCTCTTC) and *Nt.BbvCI* (CCTCAGC), and then labeled with a fluorescent-dUTP nucleotide analog using Taq polymerase. Nicks were ligated with Taq DNA ligase and the backbone of the labeled DNA was stained using the intercalating dye, YOYO-1. The nicked and labeled DNA was then loaded onto an IrysChip for imaging automatically on the Irys system (BioNano Genomics). BNG molecules were filtered with a minimum length of 150 kb and 8 minimum labels. A *p*-value threshold for the BNG assembler was set to a minimum of $2.6e-9$.

Molecules were assembled with BioNano Pipeline Version 2884 and RefAligner Version 2816.³⁴⁹

For BioNano scaffolding, *hybridScaffold.pl* version 4618 from BioNano Genomics was used. The input assembly fasta sequence was nicked *in silico* for *Nt.BspQI* and *Nt.BbvCI* labels. Consensus Maps (CMAP) were only created for scaffolds > 20 kbp with > 5 labels. A *p*-value of $1e-10$ was used as a minimum confidence value to output initial (BNG consensus map to *in silico* cmap). The final (*in silico* cmap to final hybrid cmap) alignments and a *p*-value of $1e-13$ were used as minimum confidence value to flag chimeric/conflicting alignments and to merge alignments. Scaffolds that were not super-scaffolded were added to the output from *hybridScaffold.pl*.

The BNG scaffolding pipeline identifies potential breaks that should be made to the base assembly in the form of a chimera file, but these suggested breaks are not made without manual intervention. We did not attempt to make any of the BioNano breaks. For BioNano joins, only joins that incorporated more than one scaffold were considered.

BioNano sizes gaps but does not fill them exclusively with Ns. Rather, BioNano adds in restriction site recognition sequences within the gap according to where restriction sites were seen in the BioNano map. This results in hundreds of tiny contigs which break up the BioNano gaps into smaller fragments. For the purposes of this paper, we used the GAEMR basic stats default of using 200 as a minimum contig size, effectively ignoring these restriction sites island for calculating assembly statistics and obtaining a single gap per join.

Illumina

In order to compare the completeness of assemblies constructed with different combinations of PacBio, Dovetail, and BioNano, we collected Illumina data that was

independent of the assemblies. Illumina short-insert paired ends were generated from an independent DNA sample using TrueSeq v3.0 chemistry and sequenced on an Illumina HiSeq[®] 2000. A total of 332,236,248 reads (71.4X coverage) of length 100 nt were generated.

Transcriptome assembly

To evaluate how the transcriptome was represented in the genome assemblies, the transcriptome of 14 day old R108 roots was sequenced using Illumina's RNA-Seq protocol. The transcriptome was assembled using the Transcriptome Assembly Pipeline (BPA2.1.0).³⁵³ The BPA pipeline includes a kmer sweep assembly strategy with ABySS (using the kmer values of 50, 60, 70, 80 and 90),³⁵⁴ followed by an OLC (overlap layout consensus) assembly with CAP3³⁵⁵ to find overlaps between contigs (unitigs). Scaffolding with ABySS and gap closure were performed to obtain the final assembled transcriptome sequences.³⁵⁴ The transcripts were clustered at 98% sequence identity using the CD-HIT-EST software.¹²⁵ Finally, the set of transcript sequences were filtered by length (minimum length of 100bp). An additional filtering step using ESTScan³⁵⁶ was performed to identify open reading frames using *M. truncatula* protein coding genes as a reference, yielding the final transcriptome set. Transcripts were mapped against each of the five assemblies using GMAP.³⁵⁷ Transcript hits were retained if aligning along at least 90% of their sequence with at least 90% identity.

BUSCO

Benchmarking Universal Single Copy Orthologs (BUSCO) provides a quantitative assessment of genome assemblies based on orthologs selected from OrthoDB.³¹ Assembly assessments were performed using the plant early release of BUSCO v1.1b1, which contains 956 genes that are present in at least 90% of the plant species used to assemble the database.³¹

tBLASTn searches were used to identify BUSCOs followed by Augustus gene predictions and classified into lineage specific matches using HMMER within the BUSCO package.

Read alignments

In order to assess the completeness of the assembly, PacBio filtered (minimum length of 50 and minimum quality of 75) subreads were realigned to the five assemblies using the BLASR mapper [67].³⁵⁸ All the subreads were considered for the alignment to the assemblies (-useallccs). Illumina reads were aligned to the five assemblies using the Burrows-Wheeler Aligner (BWA),²⁰⁷ version 0.7.12 with a maximum of 2 paths and sam output format.

Structural Annotation

To understand how gene sequences were affected by the assembly strategies, the MAKER-P genome annotation pipeline was used to annotate the five genome assemblies.^{183, 186, 187} All available *M. truncatula* R108 transcripts were assembled using the Trinity Assembler. All transcripts were from a single tissue, root, which is not ideal. Nevertheless, GMAP alignments to A17 indicate that the transcript assembly contains the majority of genes. Further, within the five assemblies, relative capture rates of these transcripts should not be biased by the lack of evidence transcripts from multiple tissues.

The resulting assembly was used as input for expressed sequence tag (EST) evidence for MAKER-P annotations.^{185, 359} The MAKER-P pipeline aligns the provided ESTs to the genome and creates *ab initio* gene predictions with SNAP¹⁸⁸ and Augustus^{180, 360} using evidence-based quality values. Each assembly was divided into ten chunks and processed through MAKER-P individually. Following completion of MAKER-P runs for each of the ten chunks, fasta and gff files were combined using fasta_merge and gff3_merge, respectively, included as part of the MAKER-P package.

Identification of structural rearrangements and novel sequences in R108

Each R108 PacBio-based assembly was first aligned to the A17 reference (i.e., Mt4.0) using BLAT.³⁶¹ The resulting alignments were merged, fixed (removing non-syntenic or overlapping alignment blocks) and cleaned (removing alignment blocks containing assembly gaps). BLAT Chain/Net tools were then used to obtain a single coverage best alignment net in the target genome (HM101) as well as a reciprocal-best alignment net between genomes. Finally, genome-wide synteny blocks were built for each assembly (against HM101), enabling identification of genome structural rearrangements including the chr4-8 translocation.

Based on pairwise genome comparison of R108 and A17, we obtained a raw set of novel sequences (present in R108 but absent in A17) by subtracting all aligned regions from the gap-removed assembly. Low-complexity sequences and short tandem repeats were scanned and removed using Dustmasker³⁶² and Tandem Repeat Finder³⁶³. Potential contaminant sequences (best hit in non-plant species) were filtered by BLASTing³⁵⁹ against NCBI Nucleotide (nr/nt) database. Genes with more than 50% CDS in these regions comprised the accession-specific gene set. Pfam analysis and functional enrichment were then performed on this novel gene list.³⁶⁴

List of Abbreviations:

Pb: PacBio
 Dt: Dovetail
 Bn: BioNano
 PbDt: PacBio Dovetail
 PbBn: PacBio BioNano
 PbDtBn: PacBio Dovetail BioNano
 PbBnDt: PacBio BioNano Dovetail

Availability of data and material

The R108 v1.0 assembly, sample information and the raw PacBio reads are available in Genbank (BioProject: PRJNA368719, Biosample: SAMN04571790, PacBio reads: SRS1353205, assembly MWMB000000000.1). The gene annotation (GFF3) and BioNano (BNX) files are available under DOI numbers DOI: 10.13140/RG.2.2.29595.36647 and DOI: 10.13140/RG.2.2.32950.80964, respectively. The R108 RNA-Seq data are available in the NCBI sequence read archive (SRA), under BioProject accession number [SRP077692](#).

Additional Files

Supplementary Figures and Tables. Contains Supplementary Figure S1 and Supplementary Tables S1-S5. (DOCX 110 kb) DOI 10.1186.

APPENDIX F

SOURCES AND RE-SOURCES: IMPORTANCE OF NUTRIENTS, RESOURCE
ALLOCATION, AND ECOLOGY IN MICROALGAL CULTIVATION FOR LIPID
ACCUMULATION

Fields, Matthew W., Adam Hise, Egan J. Lohman, Tisza Bell, Rob D. Gardner, Luisa Corredor, Karen Moll, Brent M. Peyton, Greg W. Characklis, and Robin Gerlach

Applied Microbiology and Technology

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☒ Published in a peer-reviewed journal

99:11

Regardless of current market conditions and availability of conventional petroleum sources, alternatives are needed to circumvent future economic and environmental impacts from continued exploration and harvesting of conventional hydrocarbons. Diatoms and green algae (microalgae) are eukaryotic photoautotrophs that can utilize inorganic carbon (*e.g.*, CO₂) as a carbon source and sunlight as an energy source, and many microalgae can store carbon and energy in the form of neutral lipids. In addition to accumulating useful precursors for biofuels and chemical feed-stocks, the use of autotrophic microorganisms can further contribute to reduced CO₂ emissions through utilization of atmospheric CO₂. Because of the inherent connection between carbon, nitrogen, and phosphorus in biological systems, macronutrient deprivation has been proven to significantly enhance lipid accumulation in different diatom and algae species. However, much work is needed to understand the link between carbon, nitrogen, and phosphorus in controlling resource allocation at different levels of biological resolution (cellular versus ecological). An improved understanding of the relationship between the effects of N, P, and micronutrient availability on carbon resource allocation (cell growth versus lipid storage) in microalgae is needed in conjunction with life cycle analysis. This Mini-Review will briefly discuss the current literature on the use of nutrient-deprivation and other conditions to control and optimize microalgal growth in the context of cell and lipid accumulation for scale-up processes.

In modern societies, petroleum-based products and fuels have strongly influenced human culture and infrastructure. For example, energy, food, and chemicals make up approximately 70% of commerce on the planet (www.eia.gov), and petroleum/hydrocarbons directly and indirectly impact these commodities. Petroleum/hydrocarbon markets have become increasingly unpredictable and cause destabilized commodity prices (*e.g.*, fuel, food). In addition, the environmental impacts from increased carbon dioxide (CO₂) without balanced CO₂ sequestration has contributed to increases in atmospheric CO₂ levels. The amount of carbon released in one year from the consumption of fossil fuels is more than 400-fold the amount of carbon that can be fixed via net global primary productivity (Dukes 2003). In order to offset the massive influx of CO₂ into the atmosphere, the utilization of renewable biofuels (*e.g.*, ethanol, butanol, H₂, CH₄, and biodiesel) is needed.

Bacillariophyta (diatoms) and *Chlorophyta* (green algae) are eukaryotic photoautotrophs that can utilize inorganic carbon (*e.g.*, CO₂) as a carbon source and sunlight as an energy source, and many microalgae can store carbon and energy in the form of neutral lipids [*e.g.*, triacylglycerides (TAGs)]. Moreover, different diatoms and algae can produce and accumulate different precursors (*e.g.*, carbohydrates, fatty acids, and pigments) that are value-added products. In addition to accumulating useful compounds for biofuels and chemical feed stocks, the use of autotrophic microorganisms can further contribute to reduced CO₂ emissions through utilization of atmospheric CO₂. For these reasons, eukaryotic photoautotrophs have been studied in the context of lipid accumulation for over 50 years and were a focus of the U.S. Department of Energy's Aquatic Species Program in the 1980s and 1990s.³⁶⁵ However, low petroleum prices

eventually eroded monetary support for alternative (and renewable) energy sources until increasing petroleum prices over the last two decades reinvigorated interest in alternatives. The advent and increased use of fracking technologies has opened up new petroleum and hydrocarbon reservoirs, and almost $\$190 \times 10^9$ was spent in the United States in 2012 to drill and “frac” for conventional hydrocarbons (www.eia.gov). However, the process of fracking increases the production rate and not the ultimate supply of hydrocarbons, and peak hydrocarbon production is predicted to occur around 2030 (www.eia.gov). Regardless of current market conditions and availability of conventional sources, alternatives are needed to circumvent future economic and environmental impacts from continued exploration and harvesting of conventional hydrocarbons.

Conservative estimates predict (assuming a lipid content of 25–30% (w/w) in microalgae) that an area equivalent to 3% of the arable cropland in the United States would be required to grow sufficient microalgae to replace 50% of the transportation fuel needs in the United States.^{366, 367} Although the interest in algal biofuels has been reinvigorated,³⁶⁸⁻³⁷⁰ significant fundamental and applied research is still needed to fully maximize algal biomass and biochemical production for biofuels and other products.

The accumulation of lipids is of substantial interest because these compounds are energy-rich biodiesel precursors.^{371, 372} Much of the reported research has focused on increasing algal lipid accumulation upon exposing cultures to a range of environmental stresses prior to harvest.^{157, 224, 372-374} Temperature variations, pH, salinity, light, osmotic, and chemical stress inducements have also been investigated with varying success.⁷³ While a stress event can increase lipid accumulation, it can also limit biomass production, but the stress scenario provides a tractable method to study and understand lipid accumulation at the laboratory-scale.⁷⁴ Because of the inherent connection between carbon (C), nitrogen (N), and phosphorus (P) in biological systems, macronutrient

deprivation has been proven to significantly enhance lipid accumulation in different diatom and algae species. While nitrogen limitation is the most commonly studied stress in green algae and diatoms; the effect of silica limitation is regularly studied in diatoms.^{67, 157, 375, 376} Light and temperature are also known stressors that can impact lipid accumulation,³⁷² and particular wavelengths have been shown to impact the rate and amount of accumulated lipid in *Chlorella*.³⁷⁷ Keeping in mind that a vast majority of living pools of C, N, and P resides in the microbial realm,³⁷⁸ much work is needed to understand the link between C, N, and P in controlling resource allocation both with respect to natural and man-made systems. In this context, a 50% replacement of transportation fuel by renewable biological sources would impose a vast nutrient demand.³⁷⁹ However, microalgal biomass/product production can be coupled to wastewater resources (*e.g.*, water, N, and P), and wastewater from agricultural, industrial, and municipal activity may provide a cost-effective source of nutrients. Agricultural and municipal wastewater can be high in N and P,³⁸⁰⁻³⁸³ and thus, there is great potential for the integration of wastewater treatment and algal biofuel/biomass production (Figure F.1). However, an improved understanding of the relationship between the effects of N, P, and micronutrient availability on cellular resource allocation (cell growth versus lipid storage) in microalgae is needed. This Mini-Review will briefly discuss the current literature on the use of nutrient-deprivation and other conditions to control and optimize microalgal culture growth in the context of cell and lipid accumulation.

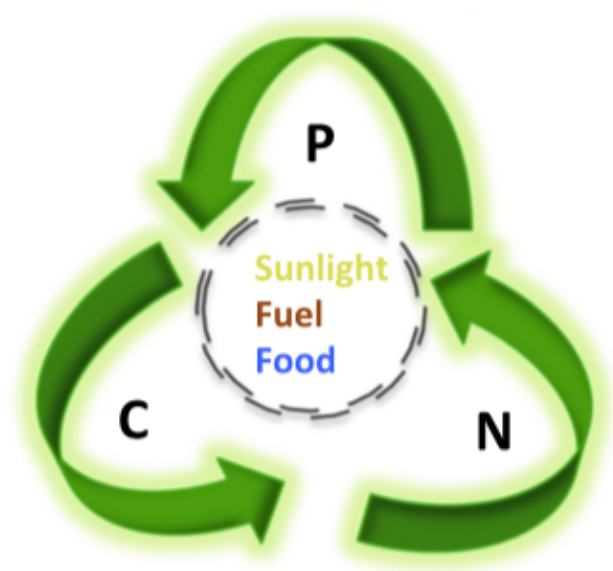


Figure F.1 The biological recycling of carbon, nitrogen, and phosphorus to harvest fuel and food linked to sunlight to reduce net consumption of N and P and net production of C.

Nutrient Dependent Lipid Accumulation

Under optimal growth conditions, (*i.e.*, adequate supply of nutrients including C, N, P and sunlight), algal biomass productivity can exceed $30 \text{ g dry weight m}^{-2} \text{ day}^{-1}$;³⁸⁴ however, the lipid content of the biomass is typically very low ($<5\% \text{ w/w}$) dependent upon species.³⁸⁴ The low lipid content is due to lipid biosynthesis being a metabolic process that is typically stimulated by stress inducement. Essentially, biomass synthesis and lipid biosynthesis compete for photosynthetic assimilation of inorganic carbon, and a fundamental metabolic switch is required to shift from biomass production to energy storage metabolism.^{385, 386} As denoted by Odum (1985), stress is a syndrome that consists of inputs and outputs, and the input is the stressor that is contrasted to the stress, or the output.³⁸⁷ Lipids (the output) are typically believed to provide a storage function within the cell that enables the organism to endure adverse environmental conditions, *i.e.*, the stressor. The output can be viewed as the cessation of cell production and the accumulation of

lipids in response to the input of unbalanced resources (e.g., N, P, and/or sunlight). It is likely that there are trade-offs in terms of biomass versus lipid accumulation depending on the different levels of perturbation (Figure F.2).

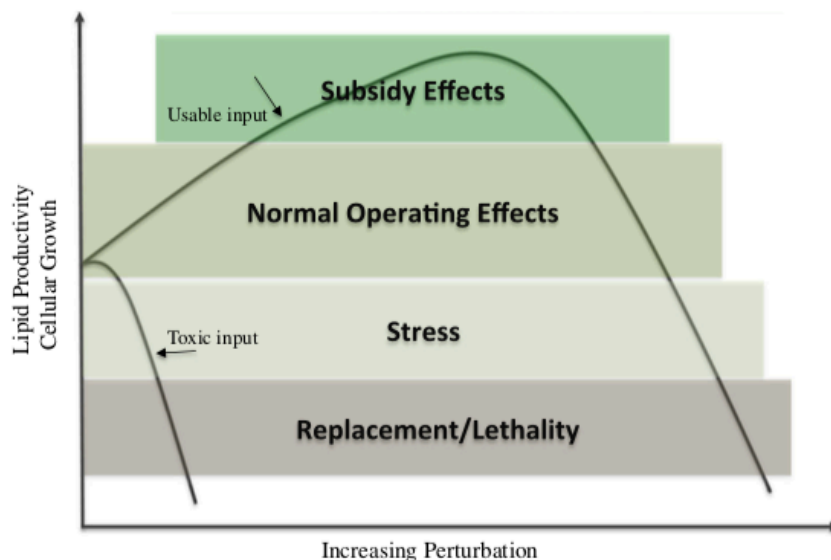


Figure F.2 Hypothetical performance curve for an increasingly perturbed (i.e., stressed) microalgal system being used to produce photoautotrophic biomass and/or lipids. Adapted from Odum et al. (1979).¹⁷⁵

Recent research has provided evidence that lipids may also act as a reservoir for specific fatty acids such as poly-unsaturated fatty acids (PUFAs).³⁸⁸ PUFAs play a key role in the structural components of cell membranes, and as antioxidants (PUFAs can counteract free radical formation during photosynthesis). As such, PUFA-rich TAGs might donate specific compounds necessary to rapidly reorganize membranes through adaptive metabolic responses to sudden changes in environmental conditions.^{389, 390} However, a recent study showed that PUFA content in lipids can negatively impact biodiesel quality based upon lipids from *Chlorella pyrenoidosa*,³⁹¹ and this

result suggests that lipid composition, and not just amounts, should be considered. In either case, lipid is an energy-rich storage compound that can be chemically transesterified to produce fatty acid methyl esters (FAME), the biological equivalent to diesel fuel (a.k.a., biodiesel). However, to maximize lipid biosynthesis, the producing organism is typically induced through environmental stress conditions.³⁷² In addition, most studies have been based upon axenic cultures with limited understanding of potential bacterial “contamination”, and thus, lipid accumulation may be different at different scales of biological resolution (discussed below).

Significant work has been done to identify and optimize stress inducement strategies that enhance lipid accumulation in microalgal species. Nutrient deprivation, specifically nitrogen depletion, is the most prevalent technique employed.³⁷² This may be due to two factors: 1) Lack of requisite nutrients such as nitrogen limits the capacity to synthesize proteins necessary for biomass production (*e.g.*, cellular division). In order to compensate, the organism must take advantage of alternative metabolic pathways for inorganic carbon fixation, such as fatty acid synthesis and hence store those *de novo* fatty acids as TAG.³⁹² 2) Photosynthesis and the electron transport chain in eukaryotic microalgae produce ATP and NADPH as energy “storage” and electron carrier metabolites, respectively.³⁹³ These metabolites are consumed during biomass production resulting in ADP and NADP⁺, which in turn are regenerated via photosystems. Under normal growth conditions, this cycle maintains a balanced ratio of the reduced and oxidized forms of these metabolites; however, when biomass production is impaired due to a lack of requisite nutrients, the pool of NADP⁺ and ADP can become depleted. This can lead to a potentially dangerous situation for the cell because photosynthesis is mainly controlled by light availability, and cannot be shut off completely. Fatty acid synthesis consumes NADPH and ATP; therefore, increased fatty acid synthesis replenishes the pool of required electron acceptors in the form of

NADP⁺, and *de novo* fatty acids are most frequently stored as lipid.³⁹⁴ Here we will review the most successful strategies involving nutrient stress to induce lipid accumulation in commonly studied microalgal species.

Nitrogen and Phosphorus

Nutrient availability is critical for cell division and intracellular metabolite cycling, and once nutrients such as N or P become depleted or limited in the medium, invariably a steady decline in cellular reproduction rate ensues. Once this occurs, the activated metabolic pathways responsible for biomass production are down-regulated and cells instead divert and deposit much of the available C into lipid.^{395, 396} There have been numerous studies to compare different N sources in the context of maximal biomass or lipid accumulation, and the results are different dependent upon the organism.³⁹⁷ *et al.* (2012) accumulated previous literature on 56 eukaryotic, photoautotrophic genera studied in the context of lipid accumulation that included (Table F.1). The authors chose *Chlorella vulgaris*, *Chlorella zofingiensis*, *Nannochloris* UTEX 1999, *Neochloris oleoabundans*, *Scenedesmus obliquus*, *Dunaliella tertiolecta*, *Isochrysis galbana*, *Phaeodactylum tricornutum*, and *Prophyridium cruentum* to conduct normalized growth and lipid accumulation studies with nitrate as the N-source.³⁹⁷ Under N-deprivation, *C. vulgaris*, *C. zofingiensis*, *N. oleoabundans*, and *S. obliquus* accumulated over 35% dry weight as TAG, and *S. obliquus* and *C. zofingiensis* had the highest TAG productivity (240 – 320 mg l⁻¹ day⁻¹) among the nine compared strains.

Table F.1 Genera of 56 eukaryotic, photoautotrophs previously studied and reported for the accumulation of lipids. Modified from Breuer et al. (2012).¹⁸⁵

<i>Amphora</i>	<i>Ankitrodesmus</i>	<i>Biddulphia</i>	<i>Botryococcus</i>
<i>Bracteacoccus</i>	<i>Chaetoceros</i>	<i>Chlamydomonas</i>	<i>Chlorella</i>
<i>Chlorococcum</i>	<i>Chroomonas</i>	<i>Cryptocodinium</i>	<i>Cryptomonas</i>
<i>Cylindrotheca</i>	<i>Dictyosphaerium</i>	<i>Dunaliella</i>	<i>Ellipsoidion</i>
<i>Emuliania</i>	<i>Enteromorpha</i>	<i>Euglena</i>	<i>Fragilaria</i>
<i>Glossomastrix</i>	<i>Gymnodinium</i>	<i>Haematococcus</i>	<i>Hantzchi</i>
<i>Hemiselmis</i>	<i>Isochrysis</i>	<i>Monallantus</i>	<i>Monodus</i>
<i>Nannochloris</i>	<i>Nannochloropsis</i>	<i>Navicula</i>	<i>Neochloris</i>
<i>Nephroselmis</i>	<i>Nitzschia</i>	<i>Ochromonas</i>	<i>Parietochloris</i>
<i>Pavlova</i>	<i>Phaeodactylum</i>	<i>Pheomonas</i>	<i>Polytoma</i>
<i>Porphyridium</i>	<i>Protophycon</i>	<i>Prototheca</i>	<i>Rhodomonas</i>
<i>Rhodorus</i>	<i>Scenedesmus</i>	<i>Scrippsiella</i>	<i>Selenastrum</i>
<i>Skeletonema</i>	<i>Stichococcus</i>	<i>Tetraselmis</i>	<i>Thalassiosira</i>
<i>Ulothrix</i>	<i>Volvox</i>		

When the model Chlorophyte *Chlamydomonas reinhardtii* was cultivated under N limitation, an increase in lipid was also observed. Interestingly, fully saturated C₁₆ fatty acids were the most abundantly synthesized compounds, whereas polyunsaturated C₁₈ fatty acids remained relatively unchanged in this organism under the tested conditions¹⁵⁷ While nitrate supported increased biomass compared to ammonium in *Monoraphidium* sp. SB2,³⁹⁸ *Chlorococcum ellipsoideum* exhibited elevated lipid levels with urea compared to nitrate.³⁹⁹ A different *Scenedesmus* strain (sp. R-16) was shown to have the highest lipid accumulation with nitrate compared to urea, peptone, or yeast extract.⁴⁰⁰ To date, nitrate is a commonly studied N source used to understand nutrient deprivation to induce lipid accumulation; however, different N sources have different effects dependent upon the organism. This is most likely a consequence of typical habitat for the organism as well as long-term life history that is common for the respective species. As the need for nutrient recycling becomes more evident, different types and mixtures of nutrients (e.g., human, agriculture, industrial) must continue to be evaluated. For example, two recent

studies investigated the ability of *Chlamydomonas polypyrenoideum* and *Chlorella pyrenoidosa* to grow and accumulate lipids during cultivation on dairy wastewater,^{149, 401} and we recently grew a green alga isolated from storage ponds of coal-bed water that produced lipids under nutrient deprivation. Nitrogen deprivation was shown to induce lipid accumulation in the wastewater isolates, *Scenedesmus* sp. 131 and *Monoraphidium* sp. 92 with ammonium, nitrate, or urea⁴⁰² or nitrate depletion in *Skeletonema marinoi*.⁴⁰³ Interestingly, *Ettlia oleoabundans* initiated lipid accumulation in response to increased temperature before nitrate was completely depleted.⁴⁰⁴ These results suggest that different combinations of potential stressors could impact lipid accumulation in different ways.

In addition to N, P starvation to induce lipid accumulation in microalgae has been studied as a sole stress or in combination with N-limitation. In general, greater lipid accumulation due to N deprivation has been observed compared to P deprivation as reported for various *Chlorella* species.^{405, 406} When the marine diatom *Phaeodactylum tricornutum* was grown under N and P limitation, an increase in lipid accumulation was noticed in all limiting conditions.^{67, 407} However, cultures of *P. tricornutum* that were limited exclusively in N showed a more significant increase in TAG than cultures that were limited solely in P. The combined limitation of both N and P resulted in the highest lipid concentrations in *P. tricornutum*.^{67, 74} Given the commonly accepted N:P ratio of 16:1 in microalgal biomass,⁴⁰⁸ the *P. tricornutum* work demonstrated that the external N:P ratio was 27 and the cellular N:P ratio was between 8:1 and 9:1 when lipid accumulation was observed.²²⁴

Both N- and P-deprivation result in cell cycle cessation, but the relative lipid accumulation response is different, and this observation is most likely a consequence of cellular resource allocation (e.g., protein/chlorophyll vs. nucleotides). Based upon results in *P. tricornutum*, we

observed a five-fold greater increase in specific fluorescence of Nile Red, a commonly used indicator of lipid accumulation,¹¹⁵ when cells were depleted of nitrate compared to cells depleted of phosphate. In addition, re-supplementation of N or P promoted cellular growth, cessation of lipid accumulation, and increased lipid consumption in *P. tricornutum*.

Carbon

It is important to keep in mind that when comparing different nutrient-deprived states, carbon above all else is absolutely required for lipid biosynthesis.⁴⁰⁹⁻⁴¹¹ Without carbon, independent of nutrient deprivation, biomass or lipid biosynthesis is impossible. Therefore, the most successful reports of lipid induction techniques in microalgal lipid production typically involve elevated concentrations of inorganic carbon in tandem with N and/or P limitation.^{107, 115, 412} These strategies often employ a CO₂ sparge to increase dissolved CO₂ above atmospheric concentrations, or addition of soluble inorganic carbon during inoculation or just prior to nutrient depletion.^{107, 115} It should be kept in mind that the addition of soluble inorganic carbon (*e.g.*, bicarbonate) can also affect pH and osmolarity. The addition of large amounts of dissolved inorganic carbon via a CO₂ gas sparge can contribute significantly to the production cost in an algal biorefinery (*e.g.*, Liu et al. 2013), and alternative methods to gaseous CO₂-based carbon supply should be considered in conjunction with pH control. Gardner et al. (2011, 2012) demonstrated that the dosage of small amounts of bicarbonate, solely or in combination with a CO₂ sparge, can achieve similar algal growth and lipid production yields compared to continuous CO₂ sparging.^{107, 115} The use of bicarbonate addition, versus CO₂ sparging, could result in significantly lower equipment costs. In either case, elevated concentrations of C, combined with N or other nutrient deprivation has been shown to induce lipid accumulation in virtually every microalgal species tested. However, an improved understanding of cellular and population

responses to not only the respective concentrations but the ratios of macronutrients (*e.g.*, C, N, and P) will improve resource utilization and promote efficient, cost-effective processes.

Silicon Limitation

Reports on silicon limitation have revealed that both marine and freshwater diatoms will accumulate lipid under Si-limiting conditions (Sharma et al, 2012),⁴¹² and diatoms possess immense potential as contributors to biodiesel production. When faced with Si-limitations, most diatoms appear to direct carbon storage towards lipid,⁴¹³ albeit the response is dependent on the degree of Si content in the cell wall. Diatoms incorporate biologically available Si as monomeric or dimeric silicic acid into silicious cell walls (frustules) and require approximately 7% of the energy expenditure required for polysaccharide cell wall formation characteristic of green algae^{59, 414, 415} Diatoms produce comparatively less cellular starch, such that fixed carbon has increased potential to be allocated to lipid accumulation.^{115, 416, 413, 417} In fact, diatom cells can accumulate enough TAG to cause the frustules to break under silica deplete conditions,⁵⁹ potentially reducing the need for energy intensive procedures associated with lipid extraction in green algae.

Numerous studies have shown increased lipid accumulation when diatoms are cultured in silica deplete media.⁴¹⁸⁻⁴²¹ However, the majority of these studies were performed on marine diatoms (*e.g.*, *Cylindrotheca spp.*, *Thalassiosira pseudonana*, and *Phaeodactylum tricornutum*) grown in media containing comparatively lower silica concentrations.^{413, 421, 422} The results of Moll et al (2014) indicate that increasing the silica concentration will increase cell numbers, which is vital for improving algal biodiesel productivity in terms of increased biomass.⁴⁹ Therefore, while research on marine diatoms for use in biofuel applications may be advantageous for use in large-scale raceway ponds due to the ability to tolerate saline environments, the actual use may be limited until conditions are optimized for diatom cell growth and lipid accumulation.

While silica limitation is known to increase lipid accumulation, combined with other physiological stresses, lipid accumulation may be enhanced. A recent study investigated the effect of coincident silica and nitrate limitation and HCO_3^- addition to promote lipid accumulation in a freshwater diatom. Moll et al. (2014) observed that combined silica and nitrate limitation, as well as sodium bicarbonate addition increased lipid accumulation compared to individual stressors with or without HCO_3^- .⁴⁹ One hypothesis for this observation is the effect on the cell cycle. Olsen et al. (13) and Vaultot et al. (20) revealed that for *Thalassiosira weissflogii* and *Hymenomonas carterae*, nitrate and silica limitation resulted in halting the cell cycle at G_1 and the G_1/S and G_2/M boundaries, respectively.⁴²³ It is possible that the two combined nutrient limitations at different periods within the cell cycle may contribute to cellular stress and ultimately lead to enhanced lipid accumulation in diatoms.

Iron Limitation

As mentioned above, N, P, and C are the most important macronutrients, but Fe is the most versatile and important trace element for biochemical catalysis. Approximately 30 to 40% of the world's oceans are iron limited, and studies have investigated "iron fertilization" experiments whereby iron is added to High Nutrient Low Chlorophyll (HNLC) areas to induce phytoplankton growth and CO_2 fixation.⁴²⁴ Iron-limited conditions are thought to alter cell physiology by reducing cell volume, chlorophyll content, and photosynthetic activity, and, thus appear to impact cellular accumulation more than lipid accumulation per se. Specifically in *P. tricornutum*, the following enzymes were down regulated during iron-starvation: β -carbonic anhydrase, phosphoribulokinase (PRK), two RuBisCO enzymes and a HCO_3^- transporter, likely resulting in decreased carbon fixation and cellular growth.⁴²⁵ The results suggest that iron limitation greatly impacts cell growth and accumulation, and that approximately $10 \mu\text{mol Fe/mol C}$ is needed by

marine algae.⁴²⁶ Iron limitation has also been linked to increased rates of silicification, thus increasing cell density and cell sinking. According to Allen et al. (2008), cells grown under Fe-limited conditions fixed carbon 14 times slower compared to cells grown in iron-replete conditions.⁴²⁵ Since iron limitation can result in detrimental physiological effects, it is pertinent to determine the potential for these processes to be useful for commercial scale lipid accumulation.

Biofilm Growth

One of the most significant limitations to the economical use of algae is the high cost of harvesting and concentrating the biomass.^{153, 427-429} To date, research has been focused on microalgae in suspended phase for lipid production, and few studies have focused on the biofilm growth state. However, the biofilm growth state provides some advantages over suspended growth systems in terms of biomass accumulation and maintenance that would be beneficial for biomass harvesting and concentrating prior to processing. Algal suspensions are often between 0.02% and 0.06% total suspended solids (TSS), and significant energy is required to harvest and concentrate the cells to 5 to 25% TSS. Biofilms can range from 6 to 16% TSS,⁴²⁹ and could potentially minimize biomass-processing costs.^{427, 428, 430} In general, the available algal biofilm studies are based upon wastewater treatment, biofilm structure and development, and aquaculture applications.^{427, 428, 431-433} There is a small amount of research on biofilm systems for the production of biomass and lipids in eukaryotic photoautotrophs,¹⁵³ but very little in relation to the influence of environmental stresses.

Recently, Schnurr et al. (2013) reported biofilm growth under nutrient starvation to stimulate lipid accumulation.⁴²⁹ A semi-continuous flat-plate parallel horizontal photobioreactor system (PBR) was designed to control the bulk medium nitrogen and silicon concentrations until nutrient depletion and biofilm onset. Wastewater was used to seed biofilm growth and was later

replaced by synthetic medium and pure cultures of *Nitzschia palea* and *Scenedesmus obliquus*. Well-attached, thick algal biofilms were observed in all experiments, until N and Si levels decreased to below detection limits, resulting in detachment from the substratum. In contrast to suspended algae, the algal biofilms did not accumulate more neutral lipids when exposed to nutrient deficient conditions in these studies. Similar results were reported by Bernstein et al. (2014) who observed little lipid accumulation in mixed culture wastewater biofilms on the field scale or in laboratory-scale algal biofilm reactors seeded with a *Botryococcus* sp. (strain WC2B). Based upon these results, there appears to be fundamental differences in the way suspended cultures and biofilm cultures respond to nutrient deprivation. The exact reasons for differences between suspended and biofilm cells is unknown, but may be a consequence of altered nutrient cycling in biofilm cells due to altered carbon flow for cellular turnover and compound accumulation. A result of ‘community growth’ (*i.e.*, biofilms) may be to accumulate excess C and reducing equivalent as cells and exopolymer rather than internal storage molecules (*e.g.*, lipid). Future work is needed to discern the differences between the physiological states of biofilm and free-living cells in multiple species.

It is possible that benthic microorganisms would prove more useful for biofilm growth modes, and we have recently grown a benthic diatom in biofilm reactors that could accumulate lipids. These results suggest that the two growth modes can elicit different behaviors, and numerous research approaches and questions need to be explored to better understand the feasibility and cellular responses of microalgal biofilms for biomass and lipid accumulation.

Ecological Effects

The literature offers many examples of increased lipid production in numerous algal species cultivated as monocultures in closed photobioreactor (PBR) or open raceway systems under varying nutrient limitations. However, as demonstrated by mathematical models and field experiments, phytoplankton biodiversity can be correlated to increased productivity.⁴³⁴⁻⁴³⁶ Furthermore, in natural freshwater systems, productivity, measured as biomass, was highest when there was abundant nutrient availability.⁴³⁷ These observations underlie the challenge of needing high biomass loads to maximize overall lipid production. Obviously, productivity can fall under the guise of several metrics ranging from biomass, cell number, chlorophyll/pigments, and more recently lipid accumulation. Despite the success of increased lipid content in nutrient deprived monocultures, recent studies indicate that comparable lipid production can also be achieved in nutrient rich systems with a diverse community.

A study by Stockenreiter et al. (2012) demonstrated increased lipid production in a naturally occurring algal communities compared to that of single species cultivated in PBRs.⁴³⁸ For freshwater systems, P is the key nutrient responsible for eutrophication and can greatly alter productivity when limited.⁴³⁸⁻⁴⁴¹ Based upon observations with PBRs with deplete nutrient availability, one would expect to see substantially higher lipid content in the oligotrophic communities than the eutrophic. However, Stockenreiter et al (2012) showed a linear increase in total algal lipid content in correlation with species richness of the examined natural communities and that lipid content in natural communities did not differ significantly from 22 laboratory monocultures (1.4×10^6 pg ml⁻¹ versus 3.3×10^6 pg ml⁻¹).⁴³⁸ Although in no way conclusive, results such as these suggest comparable lipid values in nutrient replete and deplete systems and indicate

the need to further investigate the relationship between nutrient type and abundance in the context of lipid production in mixed communities.⁴⁴²

A diverse community is also more resistant to invasion from other species that could outcompete the desired algal species.^{434, 443} Higher nutrient availability may also aid in algal cultivation by making algae less susceptible to viral infection (see below). Coupled with community diversity, the relative health of each species is an important component. “Healthy” algae (*i.e.*, cells not under nutrient stress) can be more resistant to viral infection that leads to cell lysis.⁴⁴⁴ Rhodes and Martin (2010) developed a theoretical model that implicated high nutrient availability in significantly reduced viral infection, and such scenarios will be important to consider in microalgal cultivation processes.⁴⁴⁵

In contrast, despite the success of increased lipid accumulation in PBR monocultures under nutrient limiting conditions, economic assessments indicate that PBRs operating on a large scale may not be commercially viable.⁴⁴⁶ However, some have argued for hybrid systems that utilize a combination of both closed and open systems,⁴⁴⁷ or modified PBRs such as solid-state reactors.⁴⁴⁸ In addition, if ponds are not well mixed, biomass loss due to dark respiration may impact performance for some microalgae).⁴⁴⁹ The ecology of open and closed systems will have different parameters and inputs that will need to be considered in order to control and optimize ecosystem function (e.g., biomass, lipids, value compounds). Therefore, life-cycle analyses should help direct research to identify complementarity between water footprint, nutrient sources, regional light availability, process design, and targeted lipid-producing organisms.

Integrating Life-Cycle Analysis

Algal biofuels have the potential to provide a substantial fraction of United States transportation fuel while imposing a relatively small (arable) land footprint,⁴⁵⁰ and providing opportunities for reducing water and nutrient consumption relative to first generation biofuels.⁴⁵¹ The degree to which biology and engineering can contribute to these goals will; however, be a function of the entire lifecycle.⁴⁵² A circumscribed version of that lifecycle, one involving only the production cycle (distinct from the usage cycle), includes microbial growth, dewatering/drying, extraction/conversion, and energy/input recovery stages, and each stage involves a number of choices (Figure F.3).

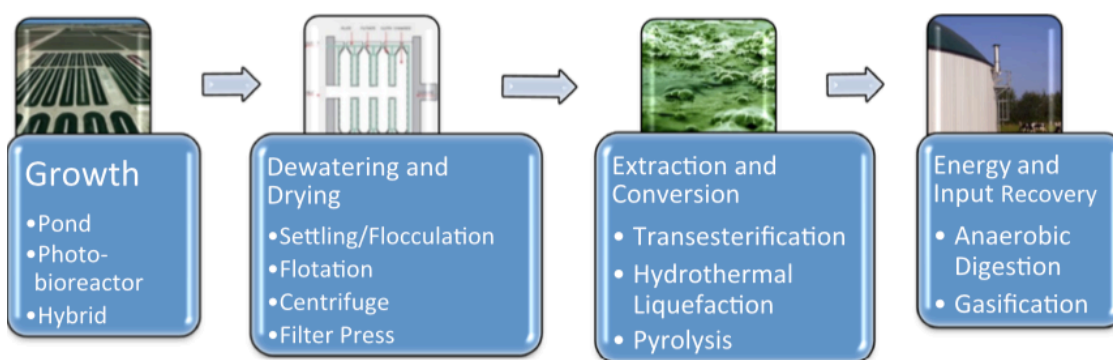


Figure F.3 Primary stages and (alternative processes) in the microalgae to fuel production process.

With respect to benefits, life cycle analysis (LCA) has promoted system optimization by highlighting processing alternatives that produce a net increase in system performance, while also avoiding environmental “burden shifting” that can be obscured when viewing the production system less holistically.⁴⁵³ In addition, choices made in the other three production stages will have implications for the growth stage, with the technology selected for extraction/conversion having particular importance.⁴⁵⁴ While each has respective strengths and weaknesses, two of the most

critical distinctions from a life-cycle perspective are (a) the degree of pre-conversion drying required⁴⁵⁵ and (b) whether the conversion process involves all of the algal biomass or only the lipid fraction.⁴⁵⁶

The dependence of transesterification processes on algal lipid content can impose extra costs in the growth stage,⁴⁵⁴ and lipid accumulation procedures typically come at the cost of algal productivity.⁴⁵⁷ As noted by Quinn et al (2013) and Chowdhury et al (2012), increasing lipid content can result in an increase in process GHG (greenhouse gas) emissions, because less residual biomass is used in a potential energy recovery stage.^{458, 459} Thus, the grid energy requirement increases proportionally with the lipid fraction. Wet extraction transesterification processes, while significantly reducing the drying energy input, typically involve solvent-based extractions that lead to concerns over solvent disposal.⁴⁶⁰ In addition, the recycling of these solvents can be challenging and energy intensive due to the high volumes and accompanying wet slurry.⁴⁶¹ Simultaneous extraction and transesterification processes (*i.e.*, “reactive extraction”) offer the potential for increased oil yields and lower process costs,⁴⁶² but the effectiveness of these processes at an industrial scale is still untested.⁴⁶³

Hydrothermal liquefaction (HTL) processes, despite greater capital expense, also reduce drying/dewatering requirements through the utilization of a wet feedstock, while converting up to 60% of the total biomass into a useable fuel product.⁴⁶⁴ HTL can result in greater fuel yields than those achieved via transesterification,⁴⁶⁵ and this technology may reduce the importance of advanced culturing methods to enhance algal lipid accumulation for biofuel production.⁴⁶⁶ However, thermochemical conversion methods such as HTL make nutrient recycling less efficient, as the nutrient rich byproducts are poorly suited for direct recycling into the growth process or

anaerobic digestion.⁴⁶⁷ In addition, N loss during the conversion process results in a substantially increased nutrient requirement in the production process.⁴⁶⁴

Life cycle analysis has been, and continues to be, successfully utilized to identify optimal algal biofuel production pathways. Ongoing refinement and application of this analytical technique can lead to advances that will guide future research toward a better understanding of the implications of many important choices, and thereby promote the development of more cost-effective and environmentally benign biofuel production processes. LCA has successfully identified synergies and tradeoffs between the growth stage and other parts of the production process, and results suggest that parallel research efforts involving both experimental research and life-cycle modeling can be effective.⁴⁶⁸

Conclusion

With the re-invigorated interest in alternative fuels, microalgae provide one option that will likely contribute to an overall plan for biomass, biochemical, and biofuel production in a more sustainable and efficient manner. Given the typical ratio of C:N:P in microalgal biomass (C₁₀₆:N₁₆:P₁), much of the research has focused on N and P (P to a lesser extent) and these two elements are linked in different ways to C through resource allocation at the cellular, population, and community levels. In addition, the supply of C either as CO₂ or bicarbonate at critical times in the growth cycle can significantly improve lipid and biomass productivity. Micronutrients also play a role in cellular responses and activity, and Si and Fe need to be further studied with respect to C:N:P ratios and the allocation of C into desired compounds (*e.g.*, lipids). Diatoms have potential for important contributions to lipid and biomass production but are less studied than the green algae. Many of the nutrient-deprived states have been studied with monocultures (or nearly

axenic) as suspended cultures, and regardless of the systems used (*e.g.*, closed reactors vs. open ponds), communities will assemble with different characteristics of stability, resiliency, and productivity. In addition, biofilms will likely develop, and may even be desired for the traits of accumulated biomass that can provide advantages for harvesting.

Moreover, while not directly covered in this mini-review, other resources/conditions will affect the cultivation of microalgae and include water, climate (*e.g.*, light and temperature), land, and location (*i.e.*, geography). Water will be essential for any biological process, and the re-cycle of water will be crucial as many parts of the globe become increasingly stressed for potable water. Light is obviously an important parameter for phototrophs, and is inherently related to temperature as the need for light energy and heat-regulation scale at different proportions. Land is an essential commodity whether bioreactors or ponds are used and should not compete with agricultural needs. The location of growth and processing facilities are crucial aspects to be considered via LCA both for economic implications as well as the biology/ecology (*e.g.*, biogeography) that can differ from region to region. Therefore, targeted science and engineering research is needed to better inform life-cycle analyses and process design to maximize productivity, efficiency, and cost-ratios.

APPENDIX G

DIRECT MEASUREMENT AND CHARACTERIZATION OF ACTIVE PHOTOSYNTHESIS
ZONES INSIDE WASTEWATER REMEDIATING AND POTENTIAL BIOFUEL
PRODUCING MICROALGAL BIOFILMS

Hans C. Bernstein, Maureen Kesaano, Karen Moll, Terence Smith, Robin Gerlach, Ross P. Carlson, Charles D. Miller, Brent M. Peyton, Keith E. Cooksey, Robert D. Gardner, Ronald C. Sims

Bioresource Technology

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-reviewed journal
- ☐ Accepted by a peer-reviewed journal
- ☒ Published in a peer-reviewed journal

Microalgal biofilm based technologies are of keen interest due to their high biomass concentrations and ability to utilize light and CO₂. While photoautotrophic biofilms have long been used for wastewater remediation, biofuel production represents a relatively new and under-represented focus area. However, the direct measurement and characterization of fundamental parameters required for industrial control are challenging due to biofilm heterogeneity. This study evaluated oxygenic photosynthesis and respiration on two distinct microalgal biofilms cultured using a novel rotating algal biofilm reactor operated at field- and laboratory-scales. Clear differences in oxygenic-photosynthesis and respiration were observed based on different culturing conditions, microalgal composition, light intensity and nitrogen availability. The cultures were also evaluated as potential biofuel synthesis strategies. Nitrogen depletion was not found to have the same effect on lipid accumulation compared to traditional planktonic microalgal studies. Physiological characterizations of these microalgal biofilms identify fundamental parameters needed to understand and control process optimization.

Key words: Microalgae; Biofilm; Biofuel; Wastewater Remediation; Photosynthesis

Photoautotrophic microorganisms are used as biotechnology platforms for many applications including biofuel production, wastewater remediation, carbon sequestration, and agriculture.⁴⁶⁹⁻⁴⁷¹ Of these, microalgal biofuel production has been identified as especially promising due to its potential for sustainable supplementation or replacement of fossil fuels.^{5, 372} Traditionally microalgae biotechnologies have focused on suspended, planktonic, culturing methodologies designed to facilitate photo-production; the capture and conversion of energy from photons into chemical energy stored in extractable biomolecules (*e.g.*, lipids). This study focuses on characterization of oxygenic photosynthesis and respiration in photo-biofilm reactors, an alternative and often under-represented growth system with benefits over planktonic culture such as high cell density which facilitates harvesting and reduces water requirements.

Biofilms are matrix-enclosed microbial cells attached to biological or non-biological surfaces.⁴⁷² Photoautotrophic biofilms, composed of microalgae and/or cyanobacteria, are ubiquitous to nearly all photic aquatic environments. An important attribute of biofilms is that they both create and are functionally controlled by gradients in substrates, products and energy sources.⁴⁷³ Spatial gradients in light have been shown to directly control rates of oxygenic photosynthesis and corresponding oxygen concentrations inside biofilms.⁴⁷⁴ Oxygen gradients in biofilms are directly influenced by diffusion rates and can result in localized supersaturated concentrations (with respect to air saturation) during active oxygenic photosynthesis. The resulting high oxygen concentrations can inhibit CO₂ incorporation and subsequent photo-production of carbon storage compounds by competing as a substrate for ribulose 1,5-bisphosphate carboxylase-oxygenase (RuBisCO) activity (Falkowski and Raven, 1997; Glud et al., 1992; Kliphuis et al., 2011).⁴⁷⁵ Thus, the characterization of spatial gradients in oxygenic

photosynthesis and respiration activities is a key consideration for microalgal biofilm-based technologies.

This study employed a recently developed rotating algal biofilm reactor (RABR) that was designed, built and tested at both the laboratory scale (lab-RABR) and pilot field scale (field-RABR) (Fig. G.1).⁴²⁸ The advantage of the RABR is the ability to simultaneously facilitate algal growth, biomass concentration and dewatering. Biofilm reactors can also reduce the water and energy requirements for biomass and photo-production compared to traditional suspended culturing strategies.²²²

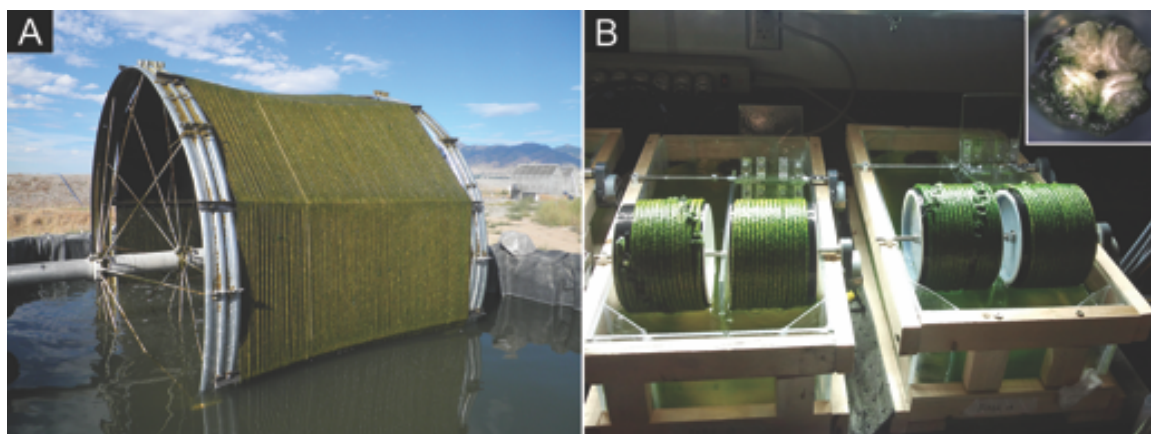


Figure G.1 Representative photographs for: (A) the field-RABR and (B) lab-RABR culturing systems designed for algal biofilm culturing (insert shows cross-sectioned excised cotton cord substratum with biofilm growth). Note the ‘top’ and ‘bottom’ biofilm orientation corresponding to the outer and inner sections of the field-RABR wheel, respectively.

The RABR and other algal-based biofilm technologies have been investigated for their potential to concurrently remediate wastewater and produce biofuel precursor molecules.^{150, 476} The RABR can facilitate efficient biomass harvesting via the reported spool harvesting technique.⁴²⁸

However, optimal biomass harvesting practices need to be determined in the context of biofilm specific physiology, such as optimal biomass areal density and biofilm thickness as it relates to active photo-production and photosynthesis zones.

The current study focuses on spatial physiological characterization of microalgal biofilms cultured through the RABR method. The specific aims of this study were to:

(1) characterize and compare two different RABR biofilms (wastewater remediating and potentially biofuel-producing) in the context of active photo-synthesis zones by directly measuring spatial gradients in steady-state oxygen and photosynthesis microprofiles, as well as, determining rates of photosynthesis and respiration processes.

(2) Characterize and compare the biofuel potential and (neutral lipid) precursor biomolecule composition in these biofilms. In addition to specific aim 2, nitrate starvation was investigated as a potential strategy for inducing lipid accumulation in the lab scale RABR biofilms.

Materials and methods

Laboratory Strains, Culturing Conditions, and Biomass Sampling.

The Chlorophyte isolate *Botryococcus* sp. strain WC-2B (hence forth referred to as WC-2B) was cultured with the 8 L lab-RABR operated in batch mode. WC-2B was isolated from an alkaline stream in Yellowstone National Park (USA), confirmed unialgal using SSU 18S rDNA and revealed 99% alignment (1,676 bp) with *Botryococcus sedeticus* UTEX 2629, which has previously been described (Senousy et al., 2004). Reactors were operated in triplicate and grown at 25°C in Bold's basal medium buffered with 25 mM 2-[N-cyclohexylamino]-ethane-sulfonic acid (CHES, pKa 9.3) and rotated at 15.3 RPM. All RABR experiments were loaded with untreated cotton cord as the biofilm-substratum (0.64 cm diameter).⁴²⁸ The lab-RABRs consisted of cords coiled onto plastic cylindrical-spools (10 cm diameter) submerged ~5 cm in the liquid medium. The lab-RABRs were cultured under custom light emitting diode (LED) banks (Box

Elder Innovations, LLC and T&L Design, Box Elder UT) programmed with LabVIEW (National Instruments Corp.) to simulate a diurnal cycle with photosynthetically active radiation (PAR) values ranging from 0-900 $\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{sec}^{-1}$ on a 14:10 L/D diel cycle following the equation:

$$I = \cos\left(\frac{\pi}{t_L} * (t - t_M)\right)^2$$

Where I is the light intensity, t_L is the total light time in minutes, t is the independent time variable, and t_M is the midpoint time corresponding to the maximum light intensity.

Medium nitrate concentrations were monitored using NitraVer 5 pillow packets (HACH). Concentrated medium (10x) and supplemental diH₂O (deionized) were added, as needed, to maintain nutrient replete conditions and offset evaporation. Culturing and sampling was performed under non-aseptic conditions (*i.e.*, open-air). Nitrate depletion was induced (after 28 days of replete culturing) by removing all liquid medium from the reactors followed by immediate replacement with Bold's basal medium without nitrate. Nitrate deplete analysis and sampling was performed 60 hr post depletion.

Biomass cell dry weights (CDW, $\text{g}_{\text{CDW}}\cdot\text{cm}^{-2}$) were obtained throughout culturing by excising a known length cotton cord and its attached biofilm, followed by biofilm removal into preweighed aluminum weigh boats. The biomass was dried at 70°C for 18 hr until the biomass weight was constant. Biomass CDWs were calculated by subtracting the dry weight of the preweighed aluminum boat from the oven dried boat with biomass and normalizing by the cylindrical surface area for a known length of cotton cord substratum.

Outdoor Culturing Conditions.

Field scale biofilms were cultured outdoors (August 10th – October 17th 2012, Logan, UT, USA) with a pilot scale RABR (field-RABR) unit constructed in accordance with previously described methods.⁴²⁸ Briefly, biofilms were grown on cotton cord (identical to lab-RABR experiments) coiled onto aluminum wheels (193 cm in diameter) which rotated (1.25 RPM) partially submerged in ~14,000 L tanks (~10,700 L liquid volume). An important difference from the lab-RABR was that the cord-substratum of the field-RABR was exposed to light and nutrients from top and bottom (discussed further below). The field-RABR was placed in a continuous flow channel of wastewater (~18.9 °C and pH~7.4) fed at ~1.25 LPM, which was drawn from the final pond of the outdoor wastewater lagoon facility (Logan, UT, USA).

Oxygen Microsensor Analysis.

Microsensor measurements were performed using Clark-type oxygen micro-electrodes with outside tip diameters of 25 µm, response time < 5 s and < 5% stirring sensitivity (Unisense, A/S).⁴⁷⁷ Amplification and sensor positioning were controlled with a microsensor multi-meter coupled with an ADC216 USB converter and a motor controlled micromanipulator. Data collection was aided by software packages, SensorTrace Pro ver.3.0.1 and Sloper ver. 3.0.3 (Unisense, A/S). Two point calibrations were performed in air saturated diH₂O ([O₂] ≈ 260 µM) and in a 1 M NaOH, 0.1 M ascorbic acid solution (anoxic standard). Calibrations were repeatedly checked in the anoxic standard and in air saturated diH₂O throughout the experiments.

Microsensor measurements were performed between 21 and 25 °C under both dark and light conditions (PAR = 700 µmol photons·m⁻²·sec⁻¹). Spatial O₂ measurements were performed in one-dimension (depth-wise) from the biofilm-air interface down towards the cotton cord substratum in 25-100 µm steps. The effective diffusion coefficient (D_e) for O₂ in the microalgal

biofilms was estimated to be $1.2 \cdot 10^{-5} \text{ cm}^2 \cdot \text{sec}^{-1}$, by assuming it to be 50% of the aqueous value corresponding to fresh water at 25 °C.⁴⁷⁸ The oxygen micro-profile and light:dark shift techniques used here have been previously described in detail.^{474, 479, 480} Briefly, Fick's law was used to calculate the total oxygen flux exported from the surface of the biofilm (net areal rate of biofilm photosynthesis or P_n) and from the photic zone inside the biofilm (net areal rate of photosynthesis of the photic zone or $P_{n,phot}$). Additionally, the light:dark shift measurements were used to estimate gross photosynthesis profiles and areal rates (P_g) which represent the total amount of oxygenic photosynthesis under the assumptions that: (i) there is an initial steady-state O_2 distribution prior to darkening, (ii) the O_2 consumption rate is identical between the light and dark time periods, (iii) the O_2 diffusion coefficient remains constant during the measurement time at each position. Detailed calculations for oxygen transport, photosynthesis, photosynthesis-coupled respiration and dark-respiration processes are included in the supplemental material for this manuscript.

Lipid Analysis.

At the time of oxygen microsensor analysis, bulk biomass was harvested from the RABRs and washed, by centrifugation and diH_2O resuspension, four times to remove medium salts. After which, the biomass was pelleted and frozen for lyophilization and lipid analysis. Extractable precursor analysis of free fatty acid, mono-, di-, tri-acyl glycerol (FFA, MAG, DAG, and TAG, respectively) was performed in accordance to the reported bead beating extraction method coupled with gas chromatography – flame ionization detection (GC-FID).¹⁵⁷ Additionally, biofuel potential, defined as total fatty acid methyl esters (FAME), produced directly from the biomass,^{70, 481} along with fatty acid profiles were determined by a previously

described method of direct *in situ* biomass transesterification and quantified with gas chromatography – mass spectroscopy.¹⁵⁷

Results and discussion

Biofilm Cultivation

Biofilms were cultured on cotton cord substratum during field and laboratory scale RABR experiments (Fig. G.1). Samples from the lab-RABR were analyzed based on nitrate replete or deplete conditions. Samples from the field-RABR were separated according to growth orientation on the substratum. The field-RABR ‘top’ and ‘bottom’ samples correspond to biofilms formed on the outer and inner section of the rotating wheel, respectively. The field-RABR top biofilms were cultured in an orientation directly exposed to ambient sun light (average daily maximum PAR = $1715 \mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{sec}^{-1}$) compared to the more shaded bottom biofilms (average daily maximum PAR = $231 \mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{sec}^{-1}$). Hence, there were four chosen biofilm sample-types analyzed and compared in this study: (i) lab-RABR biofilm that is nitrate replete, (ii) lab-RABR biofilm that is nitrate deplete (60 hr deplete culturing), (iii) field-RABR biofilm cultured on the top (outer wheel biofilm), and (iv) field-RABR biofilm cultured on the bottom (inner wheel biofilm). It is important to emphasize that the laboratory and field-RABR systems are not identical and represent two different process objectives and are intended to be compared independently of each other. However, a future goal for the

Table G.1 Measurements of areal photosynthesis rates, areal respiration rates and relevant depth scales for the laboratory- and field-RABR cultured biofilms.

Areal rates ($\mu\text{mol O}_2 \cdot \text{cm}^{-2} \cdot \text{sec}^{-1}$)	Field RABR Top Biofilm	Field RABR Bottom Biofilm	Laboratory RABR Nitrate Replete	Laboratory RABR Nitrate Deplete
Photosynthesis, P_g	^a $11.84 \cdot 10^{-4}$	^a $5.23 \cdot 10^{-4}$	^a $7.51 \cdot 10^{-4}$	^a $5.70 \cdot 10^{-4}$
Net areal rate of biofilm photosynthesis, P_n (%P_g)	$3.01 \cdot 10^{-4}$ (25.4%)	$3.55 \cdot 10^{-4}$ (67.9%)	$2.31 \cdot 10^{-4}$ (30.8%)	$2.41 \cdot 10^{-4}$ (42.3%)
Net areal rate of photic zone photosynthesis, $P_{n,phot}$ (%P_g)	$3.64 \cdot 10^{-4}$ (30.7%)	$3.96 \cdot 10^{-4}$ (75.7%)	$3.10 \cdot 10^{-4}$ (41.3%)	$2.91 \cdot 10^{-4}$ (51.1%)
Areal respiration of the biofilm, R_{light} (%P_g)	$8.83 \cdot 10^{-4}$ (74.6%)	$1.68 \cdot 10^{-4}$ (32.1%)	$5.20 \cdot 10^{-4}$ (69.2%)	$3.29 \cdot 10^{-4}$ (57.7%)
Areal respiration of the photic zone, R_{phot} (%P_g)	$8.20 \cdot 10^{-4}$ (69.3%)	$1.27 \cdot 10^{-4}$ (24.3%)	$4.41 \cdot 10^{-4}$ (58.7%)	$2.79 \cdot 10^{-4}$ (48.9%)
Respiration in the dark, R_{dark}	$0.54 \cdot 10^{-4}$	$1.11 \cdot 10^{-4}$	$0.65 \cdot 10^{-4}$	$0.74 \cdot 10^{-4}$
Depth of photic zone, L_{phot} (μm)	^b 1100 ± 200	^b 900 ± 200	^b 675 ± 25	^b 650 ± 25
Depth of oxic zone in light (μm)	^b 1750 ± 25	^b 1800 ± 25	> 2675	> 2675
Depth of oxic zone in dark (μm)	^b 700 ± 25	^b 450 ± 25	^b 850 ± 25	^b 1150 ± 25

^a Mean of 2-3 independent measurements plus or minus a range of 25% from the mean.

^b Plus or minus measurement step-size, $n = 2-3$

RABR technology is to better integrate the wastewater remediating and biofuel producing processes; hence a minimal number of comparisons based on basic biofilm physiology were made between the two systems.

The maximum specific growth rates, measured during exponential phase, were 0.09 and 0.17 day^{-1} for the laboratory and field cultured biofilms, respectively. The maximum measured biomass areal density (observed during stationary phase) were 0.36 and 0.65 $\text{g}_{\text{CDW}} \cdot \text{cm}^{-2}$ for the lab- and field-RABRs, respectively. The final biomass areal density decreased by 0.01 $\text{g}_{\text{CDW}} \cdot \text{cm}^{-2}$ 60 hr post nitrate depletion in the lab-RABR biofilms, potentially indicating minor biomass

sloughing or degradation. The measured biofilm thickness (distance from substratum to biofilm surface at late stationary phase) was approximately 1 mm for each lab-RABR biofilm (nitrate replete or deplete) and approximately 2 mm for each field-RABR biofilm (top and bottom).

Field-RABR for Wastewater Remediation

Biofilm Heterogeneity. Direct, spatially resolved measurements of steady state oxygen profiles revealed differences between the biofilms formed on the top and bottom of the field-RABR wastewater remediating system. The illuminated portions of both biofilms near the surfaces became supersaturated with O_2 , reaching concentrations over 600 μM which was approximately 3X the measured O_2 saturation of the bulk wastewater (Fig. G.2A and G.2B). Both biofilms were oxic to depth of approximately 1800 μm below the surface while illuminated (Table G.1). Steady-state oxygen profiles were also obtained after 15 min of dark conditioning (Fig. G.3C and G.3B) and the corresponding oxic-zone depths were 700 and 450 μm in the top and bottom biofilms, respectively. This is evidence for higher oxygen consumption potential in the darkened bottom oriented biofilm (discussed in more detail below).

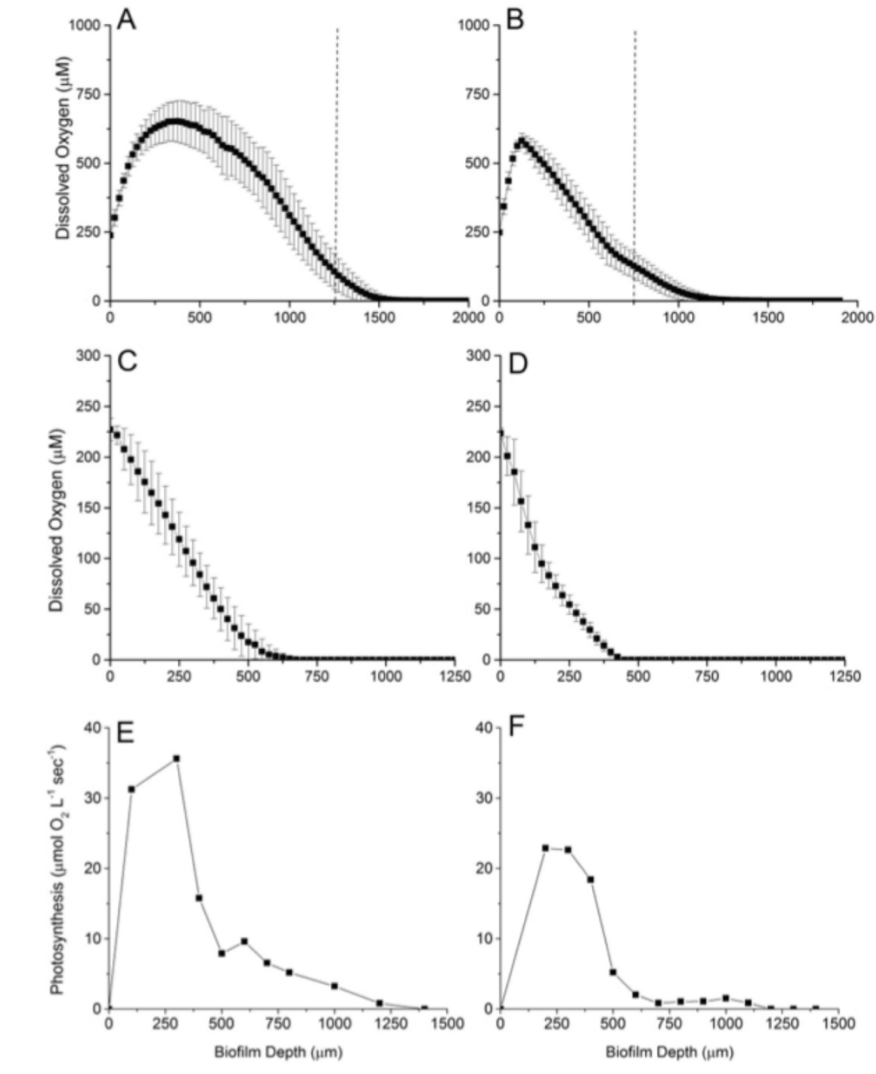


Figure G.2 Field-RABR: dissolved oxygen microprofiles measured in the light extending from the surface for biofilms grown on the (A) outer wheel surface and (B) inner wheel surface; dissolved oxygen microprofiles measured in the dark for biofilms grown on the (C) outer wheel surface and (D) inner wheel surface; and photosynthesis profiles extending from the surface for biofilms grown on the (E) outer wheel surface and (F) inner wheel surface. Note that the biofilm surface position (depth = 0) is approximated by the position at which oxygen responses were measureable (subject to $\pm 25 \mu\text{m}$ error or $\pm 100 \mu\text{m}$ error for the photosynthesis profiles where each data point is a representative gross volumetric photosynthesis rate from 2-3 replicates.) and individual data points represent the mean values from 3-4 replicate profiles in both light and dark conditions. Error bars represent plus or minus one standard deviation. Dotted lines indicate the photic-zone termination depth, estimated from the light:dark shift method. Note the scale change on the x-axis.

Oxygen gradients measured in the steady-state microprofiles show that these wastewater remediating biofilms maintain spatially varied microenvironments which may promote niche environments capable of supporting different microbial metabolic strategies. A significant portion of both biofilm samples remained anoxic during the experimental illuminated conditions (~10%) and in the dark (~50%). However, it is possible that these biofilms become fully oxic at or near peak solar irradiance during field cultivation. In the field these systems are also subject to temporal gradients in solar irradiance, temperature and nutrient flux. It is important to note that the measurements reported in this study are specific for standardized and constant incident irradiance and only represent comparative physiological potentials for these biofilms.

The field-RABR was inoculated with the native wastewater microbial flora and was composed of a complex community of environmental biofilm-forming microorganisms including phototrophs and heterotrophs. Initial 454 pyrosequence analyses indicated a high level of diversity in the field-RABR biofilms, where cyanobacteria (predominately *Oscillatoria sp.* and *Leptolyngbya sp.*) and bacterial heterotrophs accounted for significant fractions of the microbial population (Miller et al., unpublished data). However, further molecular work is required to elucidate the microbial community differences between the two biofilms with respect to their orientation of growth. It is important to reemphasize that the ‘top’ and ‘bottom’ biofilms were formed simultaneously on different sides of the same cotton cord substratum and analyzed with microsensors *ex situ* under identical conditions. Other than growth orientation, these biofilms were cultured identically and were only spatially separated by the diameter of the 0.64 cm cotton cord substratum.

Oxygenic-Photosynthesis. Direct measurements of oxygenic-photosynthesis rates quantified fundamental physiological differences in the field-RABR biofilms based only on orientation of biofilm formation (Fig. G.2E and G.2F). The measured areal rate of gross photosynthesis (P_g) in the top biofilm was ~2X greater than the bottom, signifying a much higher potential for photosynthetic electron acquisition (proportional to P_g) from the environment (Table G.1). This result was attributed to the availability of solar irradiance (PAR) during biofilm growth/formation which differed between 1715 compared to 231 $\mu\text{mol photons}\cdot\text{m}^{-2}\cdot\text{sec}^{-1}$ for the top and bottom, respectively.

The active zone of photosynthesis is defined here as the position in the biofilm where the volumetric gross photosynthesis rate [$P_g(z)$] is greater than zero and its depth assumed to be equal to the biofilm photic zone (L_{phot} , corresponding to PAR). The L_{phot} value was only slightly higher in top biofilm (Table G.1) indicating that the penetration depths of actinic light are comparable when illuminated at the same incident irradiance. The minor differences observed in L_{phot} values may translate into effective diffusion coefficient variability; however, these variances are expected to be very small based on previously reported measurements⁴⁷⁸ and were not considered here in detail. A key observation for this system is that the top oriented biofilms are capable of producing oxygen at greater than twice the rate per photon attenuated than the neighboring bottom biofilm. Additionally, the respective zones of active photosynthesis are nearly identical under standardized incident irradiance. This observation qualitatively indicates that the areal quantum yields are greater for the biofilms formed under a higher incident solar irradiance. Rigorous quantification of spatially defined quantum yields and photosynthetic efficiencies are beyond the scope of this study although the present result is consistent with established photo-physiological observations.⁴⁷⁵

Net areal photosynthesis rates were equated to the diffusive flux of oxygen transported from the biofilm surface (P_n) or the photic zone ($P_{n,phot}$) and both measurements were greater in the bottom formed biofilms compared to the top oriented samples (Table G.1). This difference is more pronounced and meaningful when interpreted as a percentage of P_g which is a proxy for the total photosynthetically derived oxygen. Net photosynthesis rates for the entire biofilm (P_n) represent 67.9% of P_g in the bottom biofilm as compared to only 25.4% in the top. These percentage differences are even greater when evaluated for $P_{n,phot}$, which includes consideration of oxygen transported to the anoxic portions in the biofilm. These results confirm that net oxygen production rates alone are not representative of the oxygenic photosynthesis potential for these samples and that the bottom orientated biofilms have the capacity to provide a greater flux of oxygen to bulk waste-water environment.

The P_n values measured for this study are only representative of steady-state reaction and diffusion processes. However, the rotating mechanism employed by the RABR alternates the biofilms between different light and fluid regimes in a periodic fashion corresponding to the submerged-liquid and ambient air surroundings. Diffusive oxygen flux was measured inside the biofilms and the steady-state microprofiles obtained on biofilms exposed to ambient air did not provide enough resolution to identify or determine the thicknesses of the diffusive boundary layer (DBL) at the surface of the biofilms. However, DBLs almost certainly were present and are not ruled out as important regulating factors in the oxygen transport processes, especially while being exposed to the liquid medium during rotation. It has been previously established that DBL thickness is a function of the velocity differential between the biofilm and bulk fluid.^{474, 482} This is an important consideration for RABR operation since the rotational speed can be optimized to reduce the effects of mass transfer limitations external to the biofilm. This highlights a future

area of characterization for the field-RABR biofilms that has the potential to enhance the biofilms productivity by minimizing mass transfer limitations.

Areal-Respiration Rates. The difference between gross and net areal photosynthesis rates provided direct measurements of photosynthesis-coupled respiration and revealed physiological distinctions between the two field-RABR biofilms. Areal photosynthesis coupled respiration rates were measured during illumination for the entire biofilm (R_{light}) and within just the photic zone (R_{phot}). Both measurements were more than 5X higher in the top formed biofilms compared to the bottom (Table G.1). Respiration rates accounted for greater percentages of P_g than the corresponding P_n values in the biofilms formed top. The opposite was true for the bottom orientated biofilms.

In contrast to the photosynthesis-coupled respiration rates, areal respiration rates in the dark (R_{dark}) were ~2X greater for the biofilms formed on the bottom biofilms as compared to the top (Table G.1). Respiration rates in corresponded directly to higher localized oxygen concentrations. This observation indicates that respiration in these biofilm consortia increases with oxygen concentration and production rate which are both functions of actinic light availability. This provides evidence of photo-respiration processes (*e.g.*, RuBisCO-oxygenase activity) acting in concert heterotrophic oxygen consumption. The bottom oriented biofilm sample has a higher capacity for light-independent heterotrophic respiration compared to the top sample which is evinced by the higher R_{dark} values.

Photosynthesis-coupled respiration is defined here to include any respiration occurring in the active zone of photosynthesis and can be advantageous to overall photo-production by lowering the localized O_2/CO_2 ratios inside the biofilm and resulting in higher selectivity for CO_2 fixation at the RuBisCO complex.^{475, 480, 483} Oxygen removal via heterotrophic or non-oxygenic

community member activity is hypothesized to be a beneficial attribute to these waste-water remediating biofilm ecosystems. Hence, the encouragement and control of localized respiration processes, independent of photo-respiration, is identified here as a potentially important design feature for RABR operation and other photosynthetic biofilm reactor technologies and should be considered for future optimization of photo-production.

The top oriented field-RABR biofilm samples showed the highest rates of gross-oxygenic photosynthesis and respiration (both R_{light} and R_{phot}). These two processes are tightly coupled inside biofilms and not considered independent from each other. In fact, it has been shown previously that photosynthesis and respiration increase concurrently with increasing irradiance in tightly controlled laboratory cultured algal biofilms.⁴⁸⁴ The differences between these closely associated wastewater remediating biofilm's capacities for photosynthesis and respiration are a result of varied solar irradiance delivered during the culturing process. The top oriented biofilms were formed with an 87% greater incident irradiance (PAR) compared to its close neighbor formed on the bottom of the cotton cord substratum. This is a practical result since it is well established that different growth environments with respect to solar irradiance availability have been shown to promote different expression levels of components comprising the light harvesting complexes, non-photosynthetic accessory pigments (*e.g.*, carotenoids) and respiration components (*e.g.*, terminal oxidases) in photosynthetic systems.⁴⁷⁵

Nitrogen Depletion in Lab-RABR Samples

Biofilm Heterogeneity. The lab-RABR biofilms, formed from the known lipid accumulating WC-2B strain, established oxygen gradients under both illuminated and dark conditions. The microprofiles revealed only subtle differences between biofilms subjected to nitrate replete and deplete conditions. Similar to the field-RABR biofilms, the illuminated

surface associated positions from both replete and deplete biofilm samples became supersaturated with O₂, reaching ~3X the measured O₂ saturation of the medium (Fig. G.3A and G.3B). During illumination, the oxic zone extended to depths greater than 2675 µm below the biofilm surface (~1675 µm into the substratum) where the flux of oxygen became very low. The WC-2B biofilms showed oxygen transport, driven by consumption, in portions of the substratum indicating that some biofilm was formed within the cotton cord pore volume. This was also observed by confocal microscopy (Supplemental Fig. S2 and S3). These lab-RABR biofilms showed a higher degree of spatial heterogeneity with respect to replicate oxygen profiles compared to the field-RABR biofilms (evident by the larger standard deviation in Fig. G.3 as compared to Fig. G.2). This increased variance between measurements taken below L_{phot} positions could result from biofilm spatial heterogeneity specific for cells attached within the cotton material. Steady-state oxygen profiles were also obtained after 15 min of dark conditioning (Fig. G.3C and G.3D). The oxic zones in the absence of light ranged from 850-1150 for the nitrate replete and deplete biofilms, respectively; indicating that the nitrogen starved biofilms had a lower potential for heterotrophic oxygen consumption (discussed in more detail below).

Oxygenic-Photosynthesis and Respiration. Direct measurements of oxygenic-photosynthesis and respiration rates quantified physiological differences in the RABR grown WC-2B biofilms cultured under nitrate replete and deplete conditions (Fig. G.3E and G.3F). Again, photosynthesis rates were measured as both net and gross production of photochemically derived oxygen at the biofilm scale. The WC-2B biofilms exhibited higher P_g values (~30%) during nitrate replete conditions indicating a greater potential for electron acquisition from the environment when not starved for nitrogen resource (Table G.1). The active zones of

photosynthesis, evaluated as the portion of the biofilm between the surface and L_{phot} , were practically indistinguishable (within 25 μm) between the two nitrate viability conditions. This measurement supports the observation that actinic light was fully attenuated by the same depth and that the oxygenic photosynthesis reaction volumes were near identical under both conditions. This observation qualitatively establishes that the WC-2B biofilms exhibit higher photosynthetic quantum yields during nitrate replete conditions.

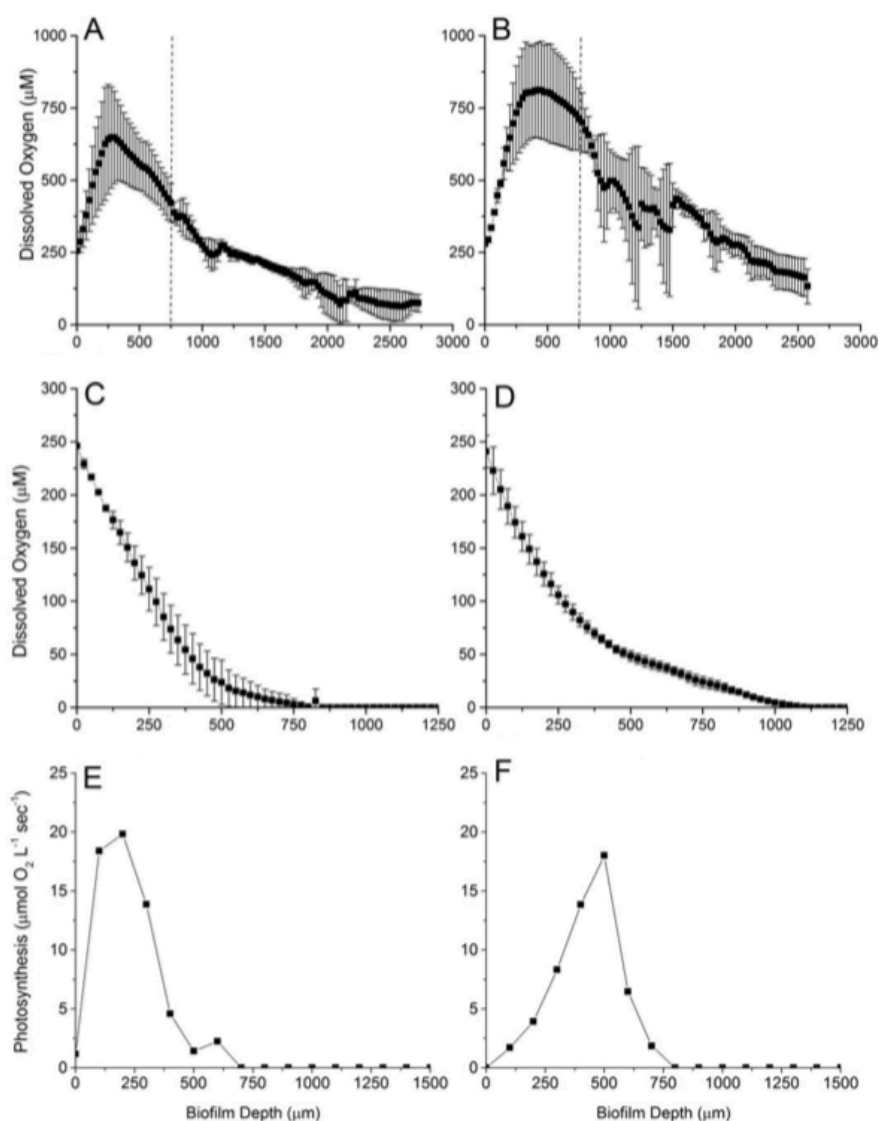


Figure G.3 lab-RABR: dissolved oxygen microprofiles measured in the light extending from the surface for biofilms grown in (A) nitrate replete and (B) nitrate deplete conditions; dissolved

oxygen microprofiles measured in the dark for biofilms grown in (C) nitrate replete and (D) nitrate deplete conditions; and photosynthesis profiles extending from the surface for biofilms grown in (E) nitrate deplete and (F) nitrate deplete. Note that the biofilm surface position (depth = 0) is approximated by the position at which oxygen responses were measurable (subject to ± 25 μm error or ± 100 μm error for the photosynthesis profiles where each data point is a representative gross volumetric photosynthesis rate from 2-3 replicates.) and individual data points represent the mean values from 3-4 replicate profiles in both light and dark conditions. Error bars represent plus or minus one standard deviation. Dotted lines indicate the photic-zone termination depth, estimated from the light:dark shift method. Note the scale change on the x-axis.

Differences in the net areal rates of photosynthesis (both P_n and $P_{n,phot}$) between the two nitrate availability conditions were not as pronounced. However, both P_n and $P_{n,phot}$ represented a greater percentage of P_g under nitrate deplete conditions. This observation is attributed to lower areal rates of photosynthetically-coupled respiration during nitrate starvation. Again, the R_{light} and R_{phot} values were measured as the difference between P_g and respective net areal photosynthesis rates. Nitrate replete conditions promoted ~20% increase in photosynthesis coupled respiration rates. The R_{dark} measurements were greater during nitrate starvation indicating a higher capacity for heterotrophic (or light independent) respiration. However, the max areal respiration rates were observed during illumination and corresponded with increased P_g . This was consistent with the observations made on the field-RABR biofilms. However, it should be noted that these lab-RABR samples are unialgal cultures and unlike the waste-water remediate system, do not represent a complex community of phototrophs and heterotrophs. Hence, respiration occurring within the lab-RABR biofilms is attributed to WC-2B physiology.

Although as a whole, there were only small differences observed in rates of photosynthesis and respiration between the two lab-RABR nitrate conditions, this data suggests two important findings. First, the results imply that nitrate depletion in the medium does not have a strong effect on the general physiology of the biofilm because only a small fraction of the biofilm (outer surface) is actively performing photochemical production under nitrogen replete

conditions. Secondly, the biofilm remained photosynthetically active under non-growth conditions hinting at the importance of maintenance energy for cell viability and the potential for nitrogen (re-)cycling. These are important observations, within the setting of algal lipid production, since nitrogen stress is a common strategy for triggering triacylglycerol accumulation in planktonic microalgal cultures.^{374, 485, 486}

The first specific aim of this study was to characterize and compare the two different RABR biofilms (lab- and field-scale) in the context of active photo-synthesis and spatial gradients in steady-state oxygen and photosynthesis. Of the physiological parameters measured for this specific aim, photosynthesis-coupled respiration is of special interest and should be considered a potent design parameter for controlling local O_2/CO_2 ratios to promote carbon fixation and subsequent photo-productivity. One potential strategy for maximizing gross photosynthesis while minimizing localized oxygen concentration would be to promote heterotrophic activity via mixed culturing techniques. Evidence for this lies in the observation that the field-RABR top-oriented biofilm community, as compared to the WC-2B lab-RABR biofilm, displayed a higher potential for electron acquisition from the environment (proportional to P_g) while channeling much greater percentages of photosynthetically derived oxygen into respiration processes. A hybridization of the wastewater remediating and biofuel production processes, may be better achieved via mixed species inoculation or ‘seeding’ with known lipid accumulating photoautotrophic community members combined with compatible heterotrophic oxygen scavengers. Consortial cooperation in microbial biofilm technology has previously been demonstrated in a number of different cell factory systems.⁴⁸⁷

Biofuel Precursor Production.

Extractable lipid fractions were recovered from all biofilm samples and analyzed by gas chromatography for assessment of biofuel properties (Table G.2). In addition, direct transesterification was performed on the lyophilized biomass to identify fatty acids and to determine total biofuel potential (extractable and non-extractable) for each biofilm-type (Table G.3). Modest increases of extractable precursor concentrations were measured in the nitrate deplete biofilms, as compared to the nitrate replete conditions. This observation was also qualitatively confirmed in microscopy images (compare Fig. S2A and S2B); where Bodipy 505/515 was used to visualize the neutral lipid precursors. The total potential FAME-weight %, representative of the total biofuel potential of the biofilm, was modestly higher for the lab-RABR biofilms that were deplete of nitrate.

The most notable observations regarding lipid production in the lab-RABR biofilms, were the differences in the total extractable weight % of lipids (sum of the FFA, MAG, DAG, and TAGs) between the nitrate replete and deplete conditions, $4.3 \pm 0.4\%$ and $7.3 \pm 0.7\%$ (w/w), respectively (Table G.2). The largest differences were observed in the DAG and TAG weight % and the respective areal concentrations. Although the WC-2B biofilms exhibited the expected, reasonable biofuel potentials; the lab-RABR production-system is not considered optimized for biofuel production. This is evident in the fact that the extractable precursors only accumulated to

Table G.2 Mean extractable biofuel precursor weight % and areal concentrations for the laboratory- and field-RABR cultured biofilms ($n = 3$ with one standard deviation error, or $n=2$ with range reported as error).

	Condition	Free fatty acid	Mono-acylglycerol	Di-acylglycerol	Tri-acylglycerol	Total extractable
Weight% (w/w)	LR-replete	2.24 ± 0.39	1.17 ± 0.21	0.76 ± 0.35	0.11 ± 0.09	4.27 ± 0.37
	LR-deplete	2.91 ± 0.46	1.48 ± 0.21	1.98 ± 0.32	0.94 ± 0.31	7.32 ± 0.70
	FR ^a	$1.49 \mid 0.50$	$0.91 \mid 0.48$	$0.34 \mid 0.12$	$0.13 \mid 0.01$	$2.86 \mid 1.09$
Areal conc. (g m^{-2})	LR-replete	0.78 ± 0.19	0.41 ± 0.10	0.26 ± 0.10	0.04 ± 0.03	1.48 ± 0.24
	LR-deplete	0.97 ± 0.22	0.50 ± 0.10	0.66 ± 0.14	0.31 ± 0.08	2.44 ± 0.38
	FR ^a	$0.98 \mid 0.33$	$0.60 \mid 0.31$	$0.22 \mid 0.08$	$0.09 \mid 0.01$	$1.89 \mid 0.72$

LR: Laboratory scale RABR.

FR: Field scale RABR.

^a Mean and range ($\mid$) for $n = 2$.

7.3 % (w/w) of the biomass, which was significantly less than planktonic cultures of WC-2B that can accumulate up to 13.9 % (w/w) of biomass as extractable precursors (7.7 % (w/w) of which is TAG) under high pH and nitrate deplete conditions (Gardner, unpublished data). This evinces that medium nitrate depletion alone may not be an effective condition for inducing TAG accumulation in microalgal biofilms, likely due to heterogeneous distributions of nutrients like nitrate caused by mass transfer limitations and the resulting distribution in microalgal activity. It should be noted that comparisons of these preliminary biofilm oil-production systems to well-mixed planktonic systems does not account for culturing times, biomass production rates or differences associated with required operating costs (*e.g.*, energy required for mixing or biomass harvesting or water input requirements).

Biofilms cultured on the field-RABR had the lowest weight percentage in both extractable precursor molecules and potential FAMES, $2.9 \pm 1.1\%$ and $5.1 \pm 1.0\%$ (w/w), respectively (Table G.2 and Table G.3). This observation coincides with the relatively high respiration rates measured in the samples (discussed earlier). The field-RABR biofilm samples are clearly not optimized for biodiesel (*i.e.*, total FAMES) production under the current culturing conditions. This could be, in part, due to colonization of non-lipid accumulating microbial community native to the wastewater (Fig. S2). However, the field-RABR exhibited higher biomass productivity ($P = \Delta g_{cdw} / \Delta \text{time}$) and total biomass areal density compared to the lab-RABR. Hence, the areal concentration of FAME transesterified from the field-RABR biofilms yielded similar values as the nitrate deplete WC-2B biofilms (Table G.3).

Table G.3 Mean FAME %, weight %, and areal concentration from the laboratory- and field-RABR cultured biofilms. Biomass was directly transesterified to determine total biofuel potential

from all fatty acid precursor molecules (extractable and non-extractable) (n=3 with one standard deviation error, or n = 2 with range reported as error).

Fatty acids	FAME%			Weight% (w/w)			Areal concentration (g m ⁻²)		
	LR-replete	LR-deplete	FR ^a	LR-replete	LR-deplete	FR ^a	LR-replete	LR-deplete	FR ^a
C12:0	N/D	0.24 ± 0.06	0.54 0.04	N/D	0.03 ± 0.01	0.03 0.01	N/D	0.01 ± 0.002	0.02 0.005
C14:0	0.39 ± 0.01	0.39 ± 0.06	1.65 0.69	0.03 ± 0.003	0.04 ± 0.01	0.08 0.02	0.01 ± 0.002	0.01 ± 0.003	0.05 0.01
C15:0	0.40 ± 0.02	0.43 ± 0.09	0.60 0.04	0.03 ± 0.003	0.05 ± 0.01	0.03 0.01	0.01 ± 0.002	0.02 ± 0.004	0.02 0.01
C16:1-3	19.85 ± 1.34	19.50 ± 0.41	22.97 2.64	1.67 ± 0.02	2.05 ± 0.03	1.17 0.36	0.58 ± 0.05	0.68 ± 0.05	0.77 0.24
C16:0	16.68 ± 1.61	18.68 ± 1.14	15.58 0.93	1.41 ± 0.23	1.96 ± 0.06	0.79 0.11	0.49 ± 0.11	0.65 ± 0.06	0.52 0.07
C18:1-3	50.15 ± 2.60	50.42 ± 1.76	42.28 2.02	4.24 ± 0.47	5.30 ± 0.34	2.15 0.52	1.47 ± 0.27	1.76 ± 0.16	1.42 0.34
C18:0	0.88 ± 0.08	0.95 ± 0.07	2.90 3.21	0.07 ± 0.01	0.10 ± 0.01	0.14 0.13	0.03 ± 0.01	0.03 ± 0.004	0.09 0.09
C20:4-5	3.35 ± 1.12	2.51 ± 0.65	5.67 0.83	0.28 ± 0.10	0.26 ± 0.06	0.29 0.10	0.10 ± 0.04	0.09 ± 0.03	0.19 0.06
C20:1-3	0.71 ± 0.13	0.64 ± 0.19	N/D	0.06 ± 0.01	0.07 ± 0.02	N/D	0.02 ± 0.01	0.02 ± 0.01	N/D
C20:0	N/D	N/D	0.18 0.37	N/D	N/D	0.01 0.02	N/D	N/D	0.01 0.01
C22:0	0.48 ± 0.08	0.51 ± 0.13	0.63 0.07	0.04 ± 0.01	0.05 ± 0.01	0.03 0.01	0.01 ± 0.004	0.02 ± 0.01	0.02 0.01
C24:0	0.28 ± 0.04	0.29 ± 0.01	0.63 0.03	0.02 ± 0.002	0.03 ± 0.000	0.03 0.005	0.01 ± 0.000	0.01 ± 0.001	0.02 0.003
C26:0	0.37 ± 0.03	0.42 ± 0.04	0.56 0.07	0.03 ± 0.003	0.04 ± 0.004	0.03 0.002	0.01 ± 0.002	0.01 ± 0.003	0.02 0.001
C28:0	N/D	0.35 ± 0.31	0.90 0.41	N/D	0.04 ± 0.03	0.04 0.01	N/D	0.01 ± 0.01	0.03 0.01
Other	6.46 ± 2.45	4.68 ± 1.30	4.90 0.05	0.54 ± 0.18	0.49 ± 0.14	0.25 0.05	0.18 ± 0.05	0.16 ± 0.04	0.16 0.03
Total	100	100	100	8.45 ± 0.61	10.51 ± 0.33	5.07 0.99	2.92 ± 0.42	3.49 ± 0.28	3.34 0.65

LR: laboratory scale RABR.

FR: field scale RABR.

N/D: Not detected.

^a Mean and range (|) for n = 2.

The second specific aim of this study was to characterize and compare the biofuel potential and (neutral lipid) precursor biomolecule composition in these biofilms. Although the current RABR systems are not considered optimized, lipid accumulation in algal biofilms is possible and reasonable if the microbial composition is constrained to known lipid producers such as the WC-2B isolate used here in the lab-RABR system. Future optimization is needed including the investigation of other industrially relevant algal strains such as *Botryococcus braunii* or *Chlorella vulgaris*. The field-RABR culturing system is a more practical and industrially scalable system compared to the lab-RABR. However, the current system is not considered viable for biodiesel production since it only accumulated $2.9 \pm 1.1\%$ and $5.1 \pm 1.0\%$ (w/w) precursor molecules and potential FAMES, respectively. Future optimization and experimentation of the field-scale system will require methodologies for enhanced control of the microbial community composition to select for better lipid accumulation. It should be noted that although biodiesel production via production of fatty acids is low in the field system, it is still a viable technique for biomass production from wastewater resource under the current conditions. Biofuel production from this system has been previously reported by using the field-RABR

derived algal biomass for acetone, butanol, and ethanol fermentation by *Clostridium saccharoperbutylacetonicum*.⁴⁸⁸

As part of the objective from specific aim 2, nitrate starvation was investigated as a potential strategy for inducing lipid accumulation in the lab-scale RABR cultures. Although a modest increase in extractable precursors was observed, nitrogen stress as implemented here by a 60 hr depletion was not identified to be as viable for “triggering” lipid accumulation in biofilms as compared to previously reported results for suspended culture studies.^{374, 489} This biofilm specific result is consistent with another previously reported study which focused on nutrient starvation (including nitrate) in cultures composed of fresh water green alga *Scenedesmus obliquus* and marine diatom *Nitzschia palea*.⁴²⁹ This previous study tested biofilm growth and lipid accumulation in algae cultured under relatively low shear in flat plate biofilm photo-reactors and reported no significant changes in lipid concentration (% dry weight) between nitrate replete and deplete conditions. This is in minor contrast to the results from the current study which observed an approximate 2-3 % w/w increase after nitrate depletion and was also qualitatively confirmed via microscopy analysis (Supplemental Fig. S2 and S3). Additionally, the Schnurr et al. study reported significant and near complete biomass sloughing post nitrate depletion which was not observed as dramatically in the lab-RABR biofilms within the 60 hr nitrate deplete phase. This could be due to the different substratum materials (*i.e.*, glass-plate compared to porous cotton cord material) and/or localized shear-stress at the biofilm surfaces. The combined results between the current and previously reported study⁴²⁹ indicate that inducing lipid accumulation via nutrient starvation may be possible but future culturing optimization is needed to evaluate the effects of known parameters associated with lipid accumulation in algal biofilms, such as nitrate and/or pH stress or chemical addition.^{374, 489, 490, 72, 386}

This manuscript explores critical photosynthetic parameters in conjunction with biofuel precursor molecule production in biofilms cultured through the novel RABR system. The lab-RABR exhibited moderate biofuel capabilities yet requires process optimization. The wastewater remediating field-RABR biofilm exhibited higher rates of photosynthesis and respiration depending on the position of biofilm formation with respect to ambient sunlight, but is not currently a viable biodiesel production platform. This study developed a methodological foundation for directly measuring photosynthetic parameters fundamental to the physiology and design of efficient photosynthetic energy harvesting platforms and establishes a benchmark for the quantitative analysis of phototrophic biofilm technologies.

Appendix G: Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.biortech.2014.01.001>.

APPENDIX H

DISSOLVED INORGANIC CARBON ENHANCED GROWTH, NUTRIENT UPTAKE, AND
LIPID ACCUMULATION IN WASTEWATER GROWN MICROALGAL BIOFILMS

Manuscript Information

Maureen Kesaano, Robert D. Gardner, Karen Moll, Ellen Lauchnor, Robin Gerlach, Brent M. Peyton, Ronald C. Sims
Bioresource Technology

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
☐ Officially submitted to a peer-reviewed journal
☐ Accepted by a peer-reviewed journal
☒ Published in a peer-reviewed journal

Microalgal biofilms grown to evaluate potential nutrient removal options for wastewaters and feedstock for biofuels production were studied to determine the influence of bicarbonate amendment on their growth, nutrient uptake capacity, and lipid accumulation after nitrogen starvation. No significant differences in growth rates, nutrient removal, or lipid accumulation were observed in the algal biofilms with or without bicarbonate amendment. The biofilms possibly did not experience carbon-limited conditions because of the large reservoir of dissolved inorganic carbon in the medium. However, an increase in photosynthetic rates was observed in algal biofilms amended with bicarbonate. The influence of bicarbonate on photosynthetic and respiration rates was especially noticeable in biofilms that experienced nitrogen stress. Medium nitrogen depletion was not a suitable stimulant for lipid production in the algal biofilms and as such, focus should be directed towards optimizing growth and biomass productivities to compensate for the low lipid yields and increase nutrient uptake.

Keywords: Microalgae, biofilms, wastewater, dissolved inorganic carbon, biofuels

Cultivation of microalgae in wastewater streams has been proposed as a means of reducing competition for freshwater sources, as an inexpensive source of nutrients, and as a biological wastewater treatment alternative.^{491, 492} Microalgae can utilize nutrients in wastewater for growth to generate considerable amounts of biomass. However, recovery of microalgae from the liquid medium is difficult and represents a substantial capital cost in suspended cultivation systems,^{369, 493} consequently there is a growing interest in attached algal growth platforms. Algal biofilm based systems such as the rotating algal biofilm reactor (RABR), algal turf scrubber (ATSTTM), revolving algal bioreactor (RAB), and Algaewheel[®] have been developed, and algal biofilm growth demonstrated in bench and pilot scale operations.^{428, 469, 494-496} However, there is still limited fundamental information on algal biofilm physiological processes and growth especially in wastewater remediation.

Widespread application of algal biofilm-based systems is also limited but can be promoted through integration of wastewater treatment with the production of valuable bioproducts from the harvested algal biomass. Algal biomass composition (*i.e.*, lipid, carbohydrate, and protein content) is influenced by the chemical composition of the medium and the environmental growth conditions (*e.g.*, temperature, pH, and light), which subsequently determines the by-products that can be synthesized. Conventionally, microalgae grown as feedstock for biofuels require a two stage process where biomass accumulation occurs under nutrient-rich conditions followed by an environmental challenge to induce secondary byproduct accumulation (*e.g.*, tri-acylglycerols as energy storage compounds).⁴⁹⁷ Nutrient starvation is typically employed as an environmental stress to stimulate lipid biosynthesis in microalgae

cultures.^{412, 498, 499} However, stimulation of lipid production in algal biofilms as a result of nutrient starvation has not been as successful as in suspended cultures.^{153, 429}

Furthermore, information on the use of other lipid inducing techniques such as chemical addition, pH stress, and temperature either independently evaluated or in combination with nutrient starvation is limited in algal biofilm studies. For example, addition of bicarbonate salts (HCO_3^-) was reported as an effective trigger for lipid production in nutrient limited suspended microalgae cultures.^{70, 490, 500, 501} The bicarbonate salts not only induce lipid production, but also provide a stable and readily available source of inorganic carbon essential for photosynthesis and microalgae growth.^{113, 374, 502} In addition, Glud et al. (1992) observed an increase in photosynthetic rates and a simultaneous reduction in respiration rates (17%) in a diatom-dominated biofilm community amended with bicarbonate.⁴⁸⁰

The potential use of bicarbonate in minimizing photorespiration is especially of interest in algal biofilms because of the high O_2/CO_2 ratios due to localized supersaturated oxygen concentration from active oxygen photosynthesis.^{153, 480} Photorespiration is a competing process to carboxylation, where ribulose-1,5-biphosphate carboxylase oxygenase (RuBisCO) acts as an oxygenase, thereby inhibiting carbon dioxide fixation and subsequently reducing photosynthetic efficiency. The study presented here evaluated the effects of adding dissolved inorganic carbon in the form of 2 mM HCO_3^- to synthetic wastewater medium to grow algal biofilms in order to:

- (1) Enhance algal biofilm growth, nutrient uptake, and lipid accumulation during nutrient deplete culturing
- (2) Increase photosynthetic rates with biofilm depth within the photic zone

Microalgal biofilm culturing and sampling

The chlorophyte isolate *Botryococcus* sp. strain WC-2B, previously described in Bernstein et al. (2014), was cultured in 8 L laboratory scale rotating algal biofilm reactors (RABRs) operated at 12 rpm and 25°C. Each reactor was comprised of two plastic cylindrical wheels (10 cm diameter) onto which 3/16 inch (diameter) untreated cotton cord was attached as the biofilm substratum. Synthetic wastewater was made to simulate typical medium strength domestic wastewater for total nitrogen (TN) and total phosphorus (TP) concentrations without a carbon source.⁵⁰³ The medium consisted of 60 mg L⁻¹ NH₄Cl, 150 mg L⁻¹ NaNO₃, 16 mg L⁻¹ Na₂HPO₄, 15 mg L⁻¹ K₂HPO₄, 4 mg L⁻¹ KH₂PO₄, 75 mg L⁻¹ MgSO₄·7H₂O, 25 mg L⁻¹ CaCl₂·H₂O, and micronutrients (8.82 mg L⁻¹ ZnSO₄·7H₂O, 1.44 mg L⁻¹ MnCl₂·4H₂O, 0.71 mg L⁻¹ MoO₃, 1.57 mg L⁻¹ CuSO₄·5H₂O, 0.49 mg L⁻¹ Co(NO₃)₂·6H₂O and 4.98 mg L⁻¹ FeSO₄).

The experimental set up consisted of four laboratory RABRs under fluorescent lights with a photosynthetically active radiation (PAR) of $227 \pm 65 \mu\text{mol m}^{-2} \text{s}^{-1}$ on a 14:10 L/D cycle. Duplicate reactors were amended with 2 mM HCO₃⁻ in the form of NaHCO₃ and another duplicate set without HCO₃⁻ amendment was cultured for comparison. The reactors were operated in sequenced batch mode with a 5 day hydraulic retention time (HRT) for a period of 18 days, after which nitrogen stress was induced for an additional 5 days by replacing all liquid medium with synthetic wastewater without a nitrogen source. For each cycle of hydraulic retention time, the reactors were drained, cleaned, and filled with fresh medium. Prior to the start of the experiment, the medium was inoculated with microalgae and the RABRs operated for 3 days (seeding period) to allow the microalgae to attach to the rope strands. As shown in Fig. 1, after the seeding period, the RABRs with the exception of the substratum (rope strands) were

covered with black polyethylene sheet to minimize microalgae growth in the liquid medium.

Culturing and sampling was performed under non-aseptic conditions (open air).

Rope samples with attached microalgae were excised for oxygen microsensor measurements, microscopy characterization, biomass dry weight measurements, and lipid analysis. Biomass cell dry weights (CDW, gdcw m^{-2}) were obtained by removing the biofilm from a known length of cord into a pre-weighed aluminum weigh boat using a flat end spatula. The biomass was dried at 70°C for 18 h until the biomass weight was constant. Biomass CDWs were calculated by subtracting the dry weight of the oven dried boat with biomass and normalizing by the total cylindrical surface area for the length of cotton cord substratum excised.

Water quality monitoring

Nitrate (NO_3^-), nitrite (NO_2^-), and orthophosphate (PO_4^{3-}) concentrations were monitored in the bulk medium and measured by ion chromatography (IC) using a Dionex IonPac AS22 carbonate eluent anion-exchange column set at a flow rate of 1.2 mL min^{-1} . IC data was analyzed by Chromeleon 7 Chromatography Data system (CDS) software. Ammonium ($\text{NH}_4^+\text{-N}$) concentrations were determined according to the 2-phenylphenol method with a BioTek PowerWave XS microplate reader (Vermont, USA) at an absorbance of 660 nm.⁵⁰⁴ The dissolved inorganic carbon (DIC) was measured on 8 mL filtered ($0.2 \mu\text{m}$ pore size filters) medium samples using a Skalar Formacs^{HT/TN} TOC/TN analyzer (model CA16, Netherlands) and Skalar LAS-160 autosampler. DIC was quantified using peak area correlation against a standard curve from a bicarbonate-carbonate mixture (Sigma Aldrich). Culture pH and optical density (OD) measurements were taken using a standard laboratory Accumet pH electrode (Fisher Scientific) and Genesys 10 UV-Model 10-S spectrophotometer (Thermo Electron Corporation), respectively.

Oxygen microsensor analysis

Clark-type oxygen microelectrodes (10 μm tip diameter; OX-10 Unisense) and specialized computer controlled hardware (Unisense) were used to analyze the reactive transport of dissolved oxygen with biofilm depth under steady-state diffusive conditions corresponding to light and dark conditions. Photosynthetic rates (coupled with photo-respiration) were estimated using the light/dark shift technique.^{474, 477} The light/dark shift measurements are valid under the following assumptions: (1) initial steady state oxygen distribution is achieved before darkening, (2) oxygen consumption rates before and after dark incubation are identical, and (3) identical diffusive fluxes are maintained during the measurement time at each position. Two point calibrations were performed for the oxic conditions (medium saturated with air) and anoxic conditions (medium sparged with nitrogen gas).

Biodiesel analysis

Biodiesel precursors *i.e.* free fatty acids (FFAs), mono-acylglycerols (MAGs), di-acylglycerols (DAGs), and tri-acylglycerols (TAGs) were extracted from dried biomass by bead beating extraction and the biodiesel potential (total FAMES) was determined by direct *in situ* transesterification according to protocols published by Lohman et al. (2013). The total FAMES and the fatty acid compositions of these FAMES were quantified using gas chromatography-mass spectroscopy (GC-MS; Agilent 6890N and 5973 Network MS). The FFAs, MAGs, DAGs, and TAGs were analyzed using gas chromatography flame ionization detection (GC-FID; Agilent 6890N).

Microalgae growth rate and yield

Microalgae successfully attached to the cotton cord and grew as a biofilm for the entire study period (Fig. S1). Curve fitting of the growth data for the entire study period showed that the 1st order growth equation provided a better description of the data compared to the zero order equation with R² values of 0.973 and 0.985 for biofilm amended with bicarbonate and those without bicarbonate respectively (supplemental data). The lag phase was minimized by the 3-day seeding period. The microalgal biofilms were in exponential growth phase from days 3 - 10 as determined from linear the portion of the log transformed growth data and the stationary phase after 10 days of growth (Fig. H.1).

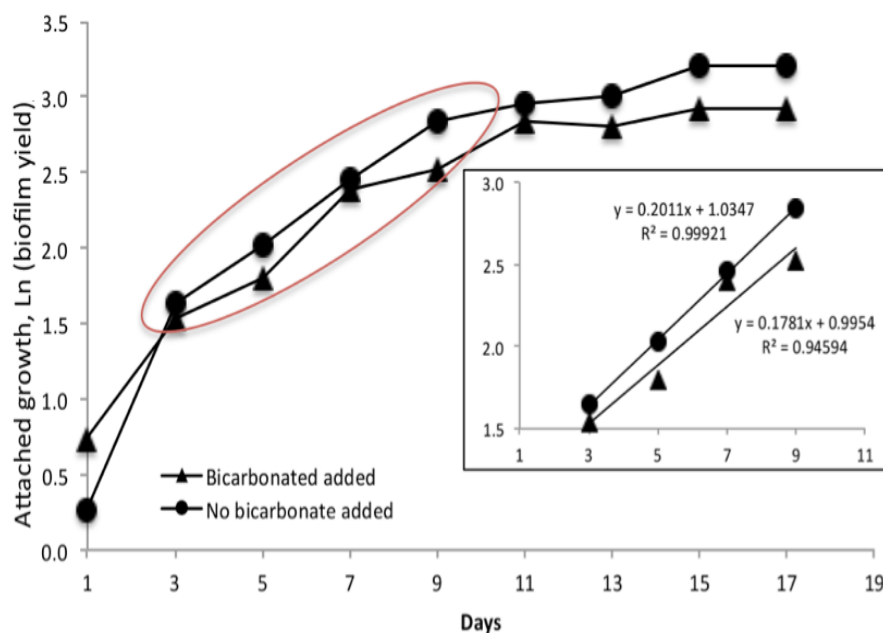


Figure H.1 Growth curve from log transformed data showing the exponential phase (day 3-10) and stationary phase (day 11 -18). Insert: Equations and R2 describing the exponential phase for both biofilms with and without bicarbonate amendment.

The maximum specific growth rates measured during the exponential phase were 0.18|0.07 (mean|range) and 0.20|0.07 day⁻¹ for algal biofilms amended with bicarbonate and the unamended control, respectively. The maximum areal biomass density measured during the stationary phase was 20.95 and 25.98 g m⁻² for biofilms with bicarbonate and biofilm samples without bicarbonate, respectively. Additionally, the biofilm production rates calculated, as the total biomass accumulated per rope surface area divided by the time taken to reach stationary phase were 1.45 and 1.79 g m⁻² day⁻¹ for biofilms with bicarbonate and biofilm without bicarbonate added, respectively.

Growth curves for the algal biofilms (attached to rope) and microalgae growth in suspension are shown in Fig. H.2A. Microalgae growth in the bulk medium was negligible over the study period indicating that covering the reactors with black plastic effectively prevented light penetration and minimized growth in suspension. There was no statistical difference observed in growth characteristics for algal biofilms amended with bicarbonate and biofilms that did not receive bicarbonate (p value of 0.4517 from t test). Although it was hypothesized that the addition of bicarbonate would increase the algal biofilm growth, this was not observed. With the 8L medium reservoirs, even the unamended algal biofilms were not carbon limited, such that bicarbonate addition did not enhance growth in this reactor system. DIC measurements remained relatively constant for each 5-day retention time with slight differences observed in the medium concentrations, with the exception of the first 3 days, (Fig. H.2B).

Removal of nitrogen and phosphorus from synthetic wastewater using algal biofilms

A basic requirement of wastewater treatment is the removal of nutrients (*i.e.*, nitrogen and phosphorus) to acceptable limits prior to discharge. Microalgae based systems promote

nutrient removal through plant uptake and subsequent harvesting of the nutrient-rich biomass from the effluent. In addition, microalgae increase the medium pH via photosynthesis thereby promoting volatilization of ammonia and possible

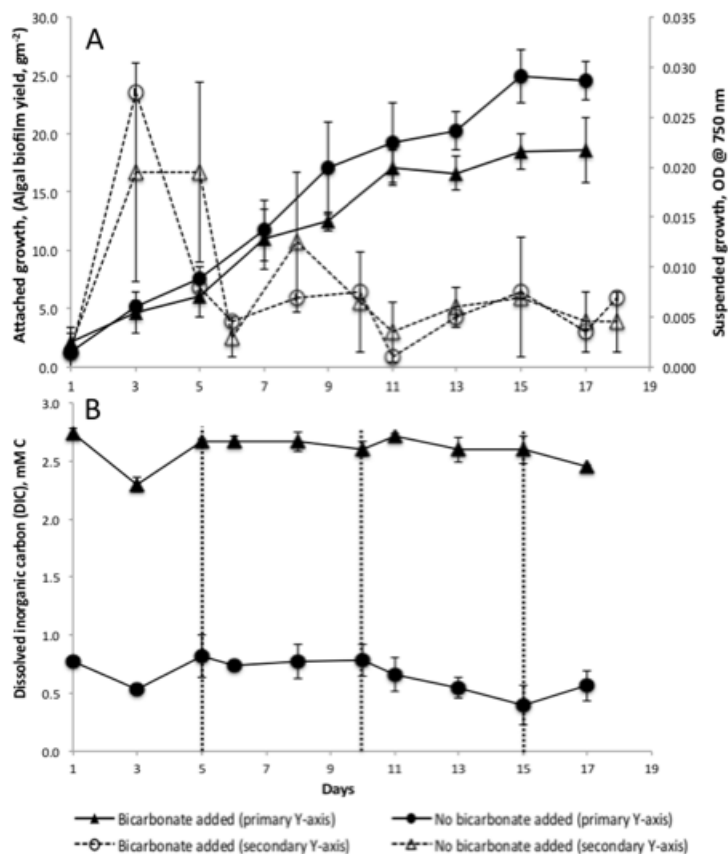


Figure H.2 Growth curves for attached and suspended microalgae (A) and dissolved inorganic carbon (DIC) concentrations (B) in laboratory-RABRs amended with bicarbonate and without bicarbonate addition. Error bars for algal biofilm yield and DIC measurements represent standard deviation (n=4). Error bars for suspended growth represent range (n=2). Vertical dotted lines represent end of 5 day hydraulic retention time.

precipitation of phosphate ions.⁴⁷⁶ It should be noted that all the RABRs were covered in black polyethylene, cleaned, and had the bulk medium replaced every 5 days to minimize algal growth in the bulk medium, which also minimized the pH increase of the medium resulting from photosynthesis. Therefore, at the measured pH of 8.5 ± 0.15 for medium amended with

bicarbonate and 7.97 ± 0.22 for medium without bicarbonate respectively, nutrient removal was attributed to the activity of the biofilms.

The synthetic wastewater was prepared with ammonia and nitrate salts as the only nitrogen sources. Initial concentrations of total nitrogen and phosphorus in the medium were approximately 40 mg-N L^{-1} and 7 mg-P L^{-1} , respectively, giving a molar N:P ratio of approximately 13:1. The measured residual total nitrogen concentrations (including NO_2^- -N) ranged from $7.95 - 19.66$ and $8.20 - 19.72 \text{ mg-N L}^{-1}$ for RABRs with and without bicarbonate amendment, respectively. Similarly, final total phosphorus concentrations ranged from $3.39 - 3.57$ and $3.35 - 3.55 \text{ mg-P L}^{-1}$ for RABRs with and without bicarbonate amendment, respectively. The lowest N and P residual concentrations were obtained during the retention time cycles corresponding to the exponential growth phase of the biofilms (Fig. H.3). Therefore, as expected, nutrient removal from the wastewater was closely linked to algal biofilm growth *i.e.*, higher removal efficiencies were obtained during the exponential growth phase of the biofilm compared to the onset of the stationary phase.

The N and P removal efficiency ranged from 27 - 74% (NO_3^- -N), 89 -100% (NH_4^+ -N), and 19 - 41% (PO_4^{3-} -P) during the experiments, with no significant difference observed between liquid samples from reactors amended with bicarbonate and those that did not receive additional dissolved inorganic carbon. Similarly, for the entire duration residual N and P concentrations followed the same trend in cultures amended with bicarbonate and those that did not receive bicarbonate (Fig. H.3). Complete uptake of ammonium ions was observed unlike nitrate ions in this study, probably due to preferential uptake of ammonia by microalgae compared to nitrate.⁴⁸¹ Microalgal cultures supplied with mixed nitrate and ammonium sources may

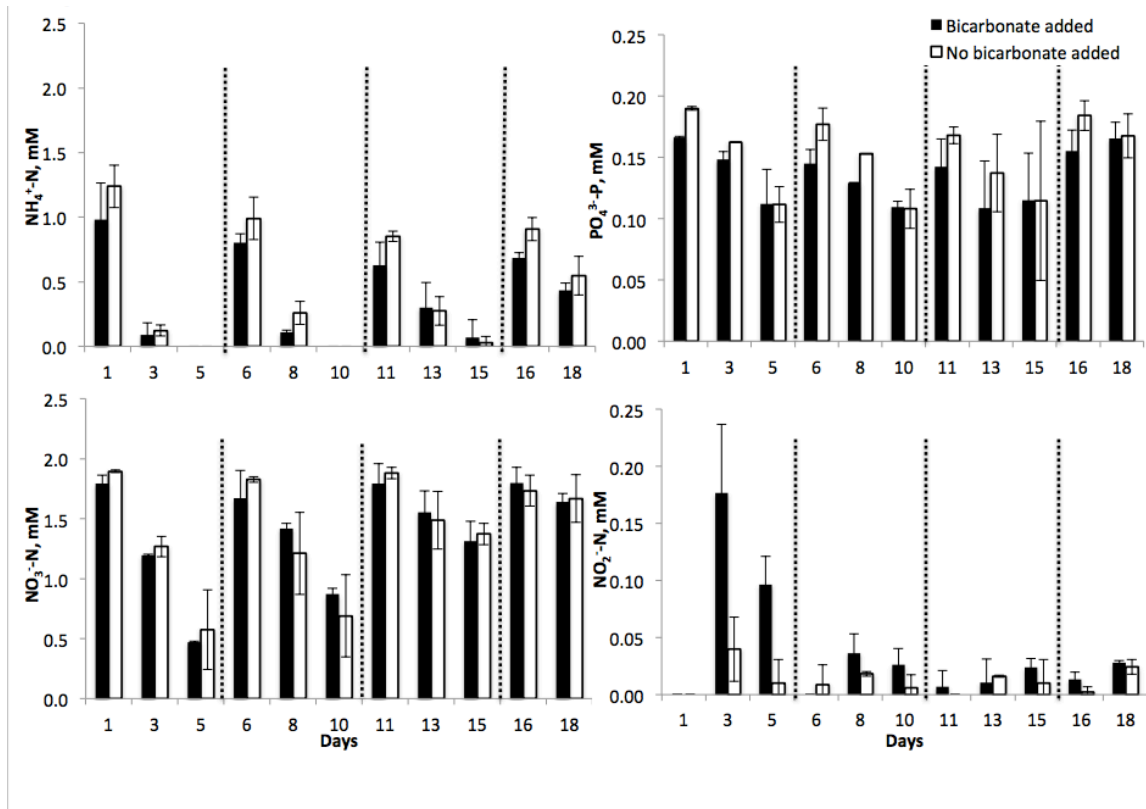


Figure H.3 Ammonium, nitrate, nitrite, and phosphate ion concentrations in medium amended with bicarbonate and without bicarbonate addition. Error bars represent range for (n=2). Vertical dotted lines represent end of 5 day hydraulic retention time.

repress $\text{NO}_3^-\text{-N}$ uptake due to feedback-inhibition, since ammonium is an end product of assimilatory nitrate reduction.⁵⁰⁵ Similar to the algal biofilm growth results, phosphate and nitrogen removal rates were not influenced by the addition of bicarbonate to the medium. Maximum nutrient removal from wastewater with algal biofilms can be attained via harvesting at the end of the exponential growth phase preferably after 8-10 days of growth using this RABR system.

The nitrite concentrations observed in solution were probably a result of incomplete nitrification of ammonia since the algal biofilms were grown in a non-aseptic oxygenated environment (Fig. H.3). An abiotic control was used to verify that the presence of $\text{NO}_2^-\text{-N}$ ions was due to biological processes (supplemental Table S1). The chemoautotrophic bacteria

involved in nitrification require a carbon source such as CO_2 or HCO_3^- , therefore the reactor with bicarbonate treatment possibly had more favorable initial conditions for the bacteria to grow, thus the higher nitrite concentrations observed (Fig. H.3). However, quasi-steady state concentrations of nitrite were eventually attained and the difference ceased to be significant later in the experiment.

Microalgal biofilm photosynthesis and coupled respiration

Oxygen microprofiles under illumination. Oxygen microprofiles were taken before and after N-deprivation was initiated, at 18 and 23 days of RABR operations. Steady state oxygen microprofiles for biofilm samples under light showed an initial increase in oxygen concentrations (compared to equilibrium with saturated air $\approx 260 \mu\text{M}$ oxygen), which peaked at a depth of $200 \pm 25 \mu\text{m}$ from the biofilm surface (biofilm/air interface) for both N-replete and N-deprived biofilms (Fig. H.4A and H.4B). Oxygen production in illuminated algal biofilms is a result of photosynthesis, and spatial gradients of light are known to affect the rate of oxygenic photosynthesis and corresponding oxygen concentrations in algal biofilms (Wieland and Kühl, 2000). Photosynthetic activity was highest in the upper layers of the biofilm and decreased with biofilm depth, possibly due to light attenuation and/or substrate diffusion limitations. Biofilms cultured under N-replete conditions had peak

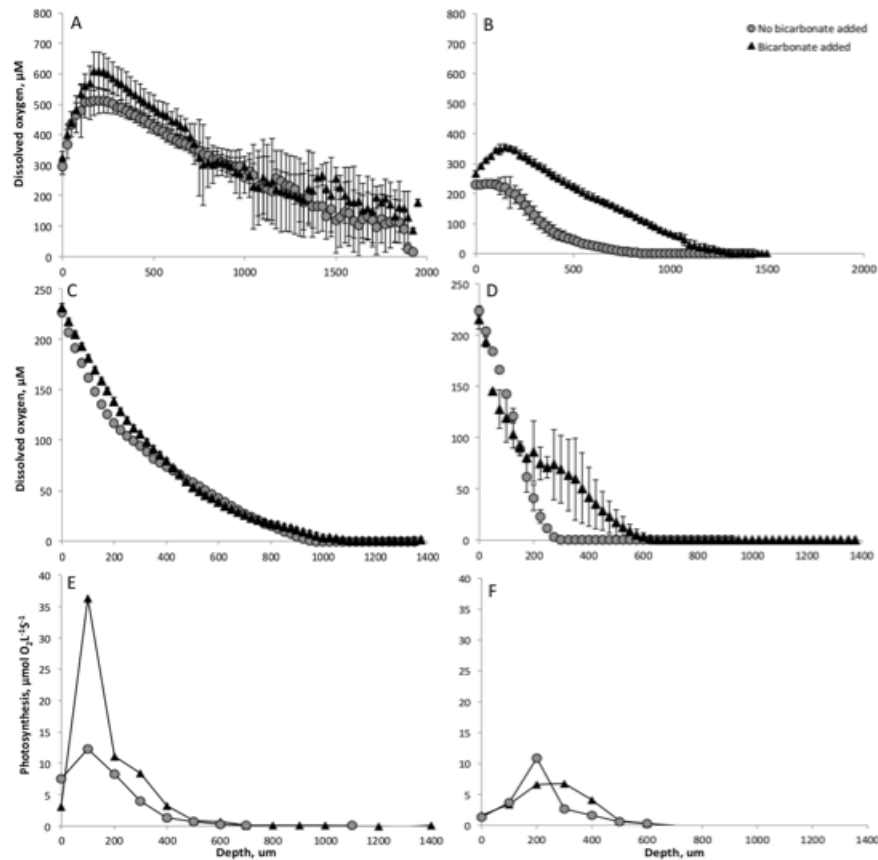


Figure H.4 Steady state oxygen microprofiles for illuminated algal biofilms under nitrogen replete (A) and nitrogen deprived (B) conditions. Error bars represent standard deviation of replicate profiles (n=3); steady state oxygen microprofiles in the dark for algal biofilms under nitrogen replete (C) and nitrogen deprived (D) conditions. Error bars represent standard deviation of replicate profiles (n=3); and representative photosynthesis profiles for algal biofilms under nitrogen replete (E) and nitrogen deprived (F) conditions. Zero depth (surface) is at the algal biofilm/air interface.

oxygen concentrations that were twice that of N-depleted biofilms (Fig. H.4A and H.4B).

Furthermore, under illumination there were no anoxic zones observed in N-replete biofilms, an indication that the oxic zone (oxygen penetration depth) extended into the cotton cord substratum (Fig. H.4A). In nitrogen replete systems, the steady state oxygen microprofiles showed no significant differences under either light or dark conditions for biofilms with or without bicarbonate (Fig. H.4A and H.4C).

On the contrary, differences in steady state oxygen microprofiles were revealed between N-deprived algal biofilms with and without bicarbonate amendment (Fig. H.4B). For example, bicarbonate amended biofilms had higher oxygen concentrations compared to biofilms that did not receive bicarbonate. This is an indication of either higher photosynthetic rates and/or reduced oxygen consumption rates due to respiration. Indeed, higher photosynthetic rates and lower areal respiration rates (in the light) were calculated for bicarbonate amended biofilm samples under N-stress (Table H.1). Additionally, anoxic zones were observed in N-deprived algal biofilms and the depth of oxygen penetration for the bicarbonate amended biofilms was 1500 μm compared to 850 μm for biofilms without bicarbonate addition (Table H.1).

Oxygen microprofiles in the dark. Oxygen is consumed by algal biofilms in the dark as a result of respiration. Assuming oxygen diffusivity is constant, the rate at which oxygen decreases (slope) is an indication of the consumption rate *i.e.*, a steeper decline in oxygen concentration indicates greater consumption and a smaller depth of oxygen penetration can be assumed to occur as a result of high heterotrophic activity.⁵⁰⁶ Steady state oxygen concentrations for biofilms in the dark decreased with depth to anoxic conditions for both N-replete and N-deprived biofilms (Fig. H.4C and H.4D). Biofilms under N-replete culturing showed a more gradual decline in oxygen concentration compared to N-

Table H.1 Measurements of photosynthetic rates, respiration rates, and relevant depth parameters for laboratory grown microalgal biofilms with and without bicarbonate amendment.

Parameter $\mu\text{mol O}_2 \cdot \text{cm}^{-2} \cdot \text{sec}^{-1}$	Bicarbonate		No bicarbonate	
	N-replete	N-deprived	N-replete	N-deprived
Gross photosynthesis, P_g	6.27E-04	2.26E-04	3.08E-04	2.02E-04
Net areal rate of biofilm	2.43E-04	9.2E-05	2.21E-04	7.43E-06
photosynthesis, P_n (% P_g)	(38.74%)	(40.76%)	(71.59 %)	(3.68%)
Net areal rate of photic zone	2.99E-04	1.36E-04	2.61E-04	5.52E-05
photosynthesis $P_{n, \text{phot}}$ (% P_g)	(47.72%)	(60.26%)	(84.87 %)	(27.33)
Areal respiration of the biofilm,	3.84E-04	1.34E-04	8.75E-05	1.94E-04
R_{light} (% P_g)	(61.26%)	(59.24%)	(28.4%)	(96.32%)
Areal respiration of the photic zone,	3.28E-04	8.97E-05	4.66E-05	1.47E-04
R_{phot} (% P_g)	(52.28%)	(39.74%)	(15.13%)	(72.67%)
Respiration in the dark, R_{dark}	0.59E-04	1.49E-04	0.74 E-04	0.98E-04
Depth of photic zone, L_{phot} , μm	1000 ± 100	600 ± 100	700 ± 100	600 ± 100
Depth of oxic zone in light, μm	>1950	1500	>1950	850
Depth of oxic zone in the dark, μm	1100 ± 25	650 ± 25	950 ± 25	300 ± 25

deprived biofilms, where steeper slopes and shorter oxygen penetration depths were observed.

This was an indication of greater potential for heterotrophic oxygen consumption in N-deprived biofilms compared to N-replete biofilms, an observation that is contrary to what was reported in Bernstein et al. (2014).¹⁵³ The current study provided a longer N-starvation period of 120 h compared to 60 h in the study by Bernstein et al. (2014), which may have promoted greater heterotrophic activity in the N-deprived biofilms.¹⁵³

Both before and after N-deprivation, biofilm samples amended with bicarbonate had greater oxygen penetration depths under dark conditions compared to biofilms that did not receive bicarbonate. Oxidic zones of $1100 \pm 25 \mu\text{m}$ and $950 \pm 25 \mu\text{m}$ in depth were estimated for biofilms amended with bicarbonate and without added bicarbonate under N-replete culturing. Similarly, oxygen penetration depths of $650 \pm 25 \mu\text{m}$ and $300 \pm 25 \mu\text{m}$ for biofilms amended with bicarbonate and without added bicarbonate during N-deprivation were observed (Table H.1). This showed that the bicarbonate amended biofilms had lower oxygen consumption in the dark compared to the biofilms without bicarbonate amendment for both nutrient conditions.

Spatial rates of photosynthesis and respiration. The gross photosynthesis profiles were generated at a spatial resolution of $100 \mu\text{m}$ vertical depth using the volumetric photosynthetic rates (*i.e.*, the rate of oxygen depletion within 3 seconds of dark incubation) determined from the light/dark shift technique. Photosynthesis occurred within a depth of $500 \mu\text{m}$ from the biofilm surface (Fig. H.4E and H.4F). Similarly, increasing rates of areal gross photosynthesis (P_g) resulted in higher areal net biofilm photosynthesis (P_n) and photic zone photosynthesis ($P_{n, \text{phot}}$), which corresponded to deeper oxidic zones (Table H.1).

However, photosynthetic rates significantly varied with both nutrient conditions and presence/absence of bicarbonate in medium. Biofilm samples under nutrient replete culturing had higher photosynthetic rates (P_g , P_n , and $P_{n, \text{phot}}$) compared to N-deprived algal biofilms indicating a greater potential for photo-productivity when nutrient replete (Fig. H.4 and Table H.1). Biofilms amended with bicarbonate also had higher photosynthetic rates (P_g , P_n , and $P_{n, \text{phot}}$) compared to the biofilms that did not receive bicarbonate for both N-replete and deprived conditions (Table H.1). The distribution of P_n and $P_{n, \text{phot}}$ as a fraction of the gross photosynthesis in the bicarbonate amended biofilms was different from that of biofilms that did not receive

bicarbonate. P_n and $P_{n,phot}$ represented a greater proportion of gross photosynthesis under N-deprived conditions for bicarbonate amended biofilms, whereas for biofilm samples that did not receive bicarbonate the reverse was observed *i.e.*, P_n and $P_{n,phot}$ represented a greater proportion of gross photosynthesis under nutrient replete conditions.

Dark respiration and photorespiration are the two basic types of respiration that occur in photosynthesizing microalgae. Dark respiration is assumed to be constant and occurs both in the light and dark whereas photorespiration is mostly active in the light and a few seconds after dark incubation.⁵⁰⁷ The dark respiration term (R_{dark}) was obtained as the slope of the initial portion of the O_2 microprofiles (linear part) in the dark. The light respiration terms (R_{light} and R_{phot}) were determined as the difference between P_g , and P_n and $P_{n,phot}$, respectively. Although, there was no clear trend observed for respiration rates (R_{light} and R_{phot}) across nutrient conditions, addition of bicarbonate to the biofilms revealed some differences. For biofilms cultured under N-replete conditions, higher areal respiration rates (R_{light} and R_{phot}) were observed in bicarbonate amended biofilms compared to biofilms that did not receive bicarbonate (Table H.1). This may have been due to the higher photosynthetic rates and subsequent increase in oxygen concentration in the biofilms amended with bicarbonate during N-replete culturing (Fig. H.4). For algal biofilms cultured under N-deprived conditions, lower R_{light} and R_{phot} were observed with added bicarbonate compared to biofilm samples without bicarbonate (Table H.1). This indicated that addition of bicarbonate reduced light respiration in N-deprived biofilms possibly due to an increased DIC supply.

Dark respiration measurements were greater for N-deprived biofilms indicating a higher capacity for heterotrophic (or light independent) respiration. The influence of bicarbonate addition on R_{dark} values varied with nutrient condition. For example, N-replete cultures had

higher R_{dark} in biofilms that did not receive bicarbonate, whereas higher R_{dark} were observed in biofilms amended with bicarbonate for N-deprived cultures (Table H.1).

Biofuel precursor production

Extractable biofuel precursor molecules (FFAs, MAGs, DAGs and TAGs) and total biofuel potential (as FAMES, i.e. extractable and non-extractable molecules) for each biofilm type, both before and after N-starvation, were measured and are presented in Table H.2 and Fig. H.5. An increase in total extractable precursor concentrations was observed in the biofilms after the 120 h N-starvation period (Table H.2). Stressed microalgae have been reported to accumulate TAG as a carbon and energy storage material.⁵⁰⁸ The sum of extractable precursors increased from 5.62% to 7.13 % (w/w) for biofilms amended with bicarbonate and 4.84% to 5.18% (w/w) for the biofilms that did not receive bicarbonate, respectively (Table H.2). Although the FFA, MAG, and DAG concentrations remained relatively constant, twice as much TAGs accumulated in the biofilms after N-starvation leading to the overall increase in total biofuel precursor molecules (Table H.2). Bicarbonate amended algal biofilms had higher weight percentage of extractable molecules.

Table H.2 Total and percent composition of extractable biofuel precursor weight (%) in laboratory grown algal biofilms with and without bicarbonate amendment.

Precursor molecules, Weight % (w/w)	Condition			
	Nutrient Replete		Nutrient deplete	
	^a Bicarbonate	^a No bicarbonate	^a Bicarbonate	^a No bicarbonate
C14 FFA	1.44 0.01	0.67 0.14	1.07 0.18	1.23 0.15
C16 FFA	1.11 0.26	1.49 0.03	0.86 0.58	1.45 0.21
C18 FFA	0.73 0.21	1.53 0.05	0.48 0.35	0.79 0.13
C16 MAG	0.11 0.05	0.18 0.01	0.11 0.09	0.13 0.03
C18 MAG	0.11 0.01	0.16 0.01	0.09 0.04	0.13 0.02
C16 DAG	0.09 0.03	0.10 0.02	0.09 0.05	0.10 0.00
C18 DAG	0.21 0.06	0.19 0.06	0.19 0.06	0.17 0.01
C16 TAG	0.49 0.41	0.17 0.08	0.58 0.41	0.35 0.15
C18 TAG	1.32 1.30	0.36 0.37	3.65 1.51	0.84 0.40
Sum of extractables				
Weight % (w/w)	5.62 1.08	4.84 0.71	7.13 0.57	5.18 0.42
Areal concentration (gm ⁻²)	1.03	1.22	1.30	1.30

^aMean and range (|) for n=2

The total FAME-weight percent and yield for N-replete and N-starved biofilms with or without bicarbonate amendment were similar (Fig. H.5A and H.5C). Although, the total FAME potential ranged from 12 – 20 % (w/w) of the biomass (Fig. H.5B), the total extractable lipids were less than 10% (w/w) (Table H.2). As previously reported by Bernstein et al. (2014), the most notable difference regarding lipid production in the RABR- grown algal biofilms was the difference in the total extractable weight percent of lipids between the N-replete and deplete conditions (Table H.2). Depletion of nitrogen and addition of dissolved inorganic carbon in the medium were not effective in stimulating substantial lipid production in the microalgal biofilms. Qualitative analysis of lipid profiles using images from CLSM showed the same result, the microalgal biofilms only showed a slight increase in lipids after N-starvation (Supplemental Fig. S4). Previous studies have attributed the inability of N-depletion in the growth medium to induce

lipid production in algal biofilms to possible nutrient re-cycling within the biofilms and resilience of algal biofilms to environmental stress.^{153, 376}

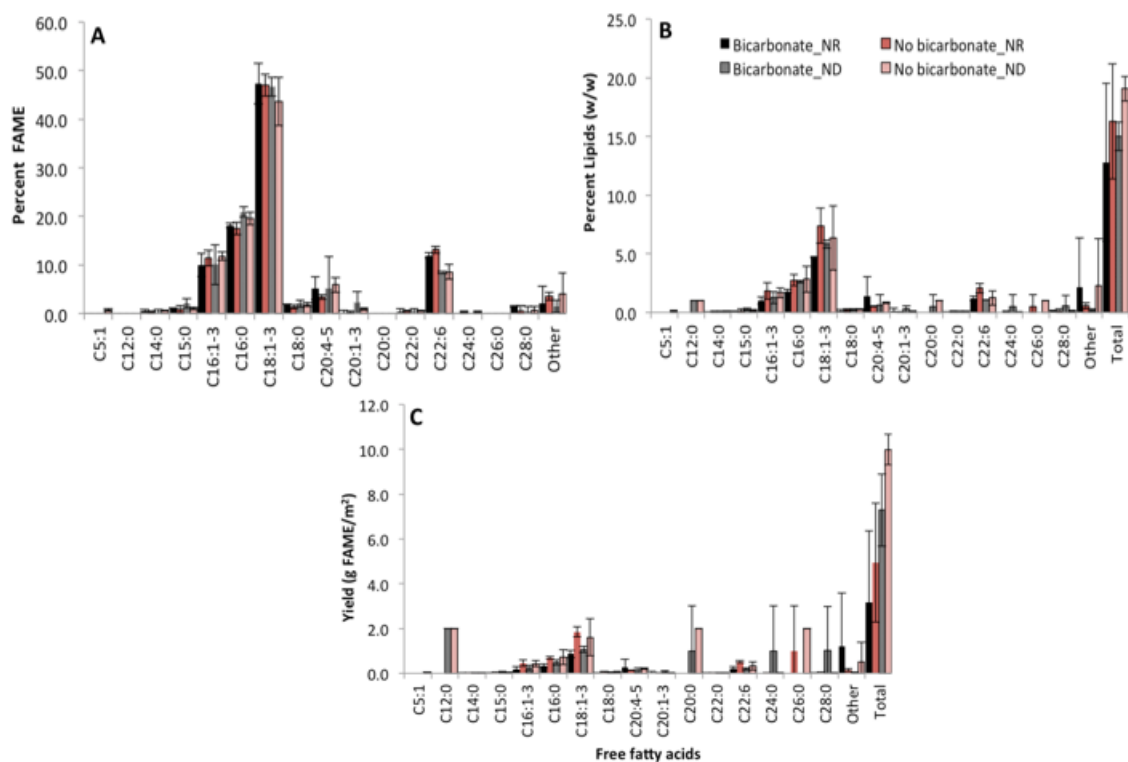


Figure H.5 Total FAMES and free fatty acid composition of the FAMES. A: Mean percent FAME (w/w), B: percent lipids (w/w), C: areal concentration (g m⁻²). Error bars represent range (n=2). ND and NR represent nitrogen deprived and replete algal biofilms, respectively.

Conclusions

For this study, there was no significant difference in algal biofilm growth, nutrient removal, and lipid accumulation between algal biofilms amended with bicarbonate and those that did not receive bicarbonate. However, an increase in photosynthesis rates was observed in algal biofilms amended with bicarbonate. The influence of bicarbonate on photosynthetic and

respiration rates was especially noticeable in biofilms that experienced nitrogen stress, as compared to biofilms in nutrient replete conditions.

Medium N-depletion may not be a suitable stimulant for lipid production in algal biofilms; rather focusing on optimizing growth, nutrient removal rates, and/or biomass productivities may be more beneficial.

Appendix H: Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.biortech.2014.12.082>. 2