

A multi-omic analysis of the photosynthetic endosymbioses of *Paramecium bursaria*

A DISSERTATION PRESENTED
BY
FINLAY MAGUIRE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY COLLEGE LONDON
LONDON, UNITED KINGDOM
DECEMBER 2015

I, FINLAY MAGUIRE, CONFIRM THAT THE WORK PRESENTED IN THIS THESIS IS MY OWN. WHERE INFORMATION HAS BEEN DERIVED FROM OTHER SOURCES, I CONFIRM THAT THIS HAS BEEN INDICATED IN THE THESIS.

© 2015 - FINLAY MAGUIRE

All rights reserved.

Author List

The following people contributed experimental work to Chapter 3: David Milner, Katie Jones, Ben Jenkins
The following people contributed experimental work to Chapter 4: David Milner, Ines Yang
The following people contributed experimental work to Chapter 5: David Milner, Scott Campbell
The following people contributed experimental work to Chapter 6: David Milner

Publications

1. Maguire *et. al.* 2014, *Complex patterns of gene fission in the eukaryotic folate biosynthesis pathway* Genome Biology and Evolution **6** 2709-2720
2. Maguire & Richards 2014, *Organelle evolution: a mosaic of 'mitochondrial' functions* Current Biology **24** (11) 518-520
3. Chambouvet *et. al.* 2015, *Cryptic infection of a broad taxonomic and geographic diversity of tadpoles by Perkinsea protists* Proceedings of the National Academy of Sciences **122** (34) E4734-E4751
4. Richards *et. al.* 2015, *Molecular diversity and distribution of marine fungi across 130 European environmental samples* Proceedings of the Royal Society B **282** 20152243
5. Chambouvet *et. al.* 2014, *Diverse molecular signatures for ribosomally 'active' Perkinsea in marine sediments* BMC Microbiology **14** (110)
6. Smith *et. al.* 2014, *A balloon-based payload for Exposing Microorganisms in the Stratosphere (E-MIST)* Gravitational and Space Research **2** (2) 70-80

A multi-omic analysis of the photosynthetic endosymbioses of *Paramecium bursaria*

ABSTRACT

The photosynthetic endosymbioses between *Paramecium bursaria* and its green algal endosymbionts (*Chlorella variabilis*, *Chlorella vulgaris*, *Micractinium reisseri* and *Coccomyxa* sp.) have long been suggested to represent nascent endosymbiotic interactions as host and endosymbiont are believed to be able to exist and reproduce separately. Understanding the molecular systems underpinning these relationships would therefore provide a model systems to understand the process of photosynthetic endosymbioses before molecular co-dependence has become fixed (leading to genomic integration). To this end, the metatranscriptome of *P. bursaria*-*M. reisseri* during lit and dark conditions was recovered using single cell methods. This necessitated the development of novel techniques to optimise the assembly and the post-assembly attribution of transcripts to their originating organism. This work represents the first de novo single cell transcriptomic analysis of a multimember eukaryotic system.

In combination with a *P. bursaria*-*C. variabilis* transcriptome and mass-spectrometry metabolomics data, this data was used to investigate metabolic function in endosymbioses. This identified potential roles for novel sugar, amino acid, and fatty acid interactions in the *M. reisseri* endosymbiosis. Additionally, *P. bursaria* SW1 was discovered to form an obligate host of *M. reisseri* SW1-ZK.

This work also reveals a putative non-functional exogenous RNA incuded RNAi system in *P. bursaria* likely related to the absence of a factor associated with uptake of RNA from host vacuoles in both *P. bursaria* transcriptomes and a partial *P. bursaria* single cell genome. An analysis of the level of potential RNAi “cross-talk” collisions with the active host transcriptome suggest that the function of an exogenous RNA induced RNAi system in the presence of a eukaryotic endosymbiont may be deleterious.

Therefore, despite discovering several barriers to the utility of these systems for endosymbiotic evolution there is still utility in their study. The “omic” resources presented here offer an important resource in guiding further analysis.

Contents

| | | |
|---------|--|----|
| 1 | INTRODUCTION | 15 |
| 1.1 | Endosymbiosis | 15 |
| 1.1.1 | What is endosymbiosis? | 15 |
| 1.1.2 | Plastid Endosymbioses | 18 |
| 1.2 | <i>Paramecium bursaria</i> | 19 |
| 1.3 | <i>Paramecium bursaria</i> - green algal endosymbioses | 24 |
| 1.4 | Conclusion | 28 |
| 2 | METHODS | 31 |
| 2.1 | Microbiology | 31 |
| 2.1.1 | Strain information | 31 |
| 2.1.2 | Media and Culture conditions | 32 |
| 2.2 | Omics | 32 |
| 2.2.1 | Genomics and Transcriptomics | 33 |
| 2.2.1.1 | DNA sequencing | 33 |
| 2.2.1.2 | Read pre-processing | 38 |
| 2.2.1.3 | Assembly | 40 |
| 2.2.1.4 | The problem with ploidy | 41 |
| 2.2.2 | Metabolomics | 41 |
| 2.3 | Machine Learning and Statistical Pattern Recognition | 43 |
| 2.3.1 | Supervised Learning | 45 |
| 2.3.1.1 | Support Vector Machines | 47 |
| 2.3.2 | Unsupervised Learning | 49 |
| 2.3.2.1 | K-means | 50 |
| 2.4 | Phylogenetics | 52 |
| 2.4.1 | Sequence sampling | 54 |
| 2.4.2 | Multiple Sequence Alignment (MSA) | 55 |
| 2.4.3 | Masking | 56 |
| 2.4.4 | Substitution model selection | 57 |
| 2.4.5 | Phylogenetic inference | 58 |
| 2.4.5.1 | Maximum likelihood | 59 |
| 2.4.5.2 | Bayesian | 60 |
| 2.5 | Informatics languages and hardware | 61 |
| 2.5.1 | Languages and Libraries | 61 |
| 2.5.2 | Hardware | 62 |
| 3 | ENDOSYMBIONT DIVERSITY | 63 |
| 3.1 | Introduction | 63 |
| 3.1.1 | Endosymbiont taxonomy and clonality | 63 |
| 3.1.2 | ITS ₂ taxonomic profiling | 64 |
| 3.1.3 | Isolation of <i>P. bursaria</i> | 65 |
| 3.2 | Aims | 66 |
| 3.3 | Methods | 66 |
| 3.3.1 | Taxonomic analysis | 66 |
| 3.3.1.1 | ITS ₂ sequencing | 66 |
| 3.3.1.2 | Phylogenetics | 67 |
| 3.3.2 | Single cell genomics | 68 |
| 3.3.2.1 | DNA extraction | 68 |
| 3.3.2.2 | Sequencing | 68 |
| 3.3.2.3 | Read pre-processing | 68 |
| 3.3.2.4 | Assembly | 69 |

| | | |
|---------|---|-----|
| 3.3.2.5 | Assembly assessment | 69 |
| 3.3.2.6 | Contig binning | 69 |
| 3.3.2.7 | Variant calling | 70 |
| 3.3.3 | Endosymbiont elimination | 70 |
| 3.4 | Results | 71 |
| 3.4.1 | ITS ₂ phylogeny | 71 |
| 3.4.2 | Single cell genomes | 73 |
| 3.4.2.1 | Sequencing and pre-processing | 73 |
| 3.4.2.2 | Assembly | 73 |
| 3.4.2.3 | Binning | 76 |
| 3.4.2.4 | Variant Calling | 79 |
| 3.4.3 | Elimination of endosymbiont | 79 |
| 3.5 | Discussion | 81 |
| 3.5.1 | CCAP 1660/12 and CCAP 1660/13 contain largely clonal <i>M. reisseri</i> symbionts | 81 |
| 3.5.2 | Reliability of culture collection | 82 |
| 3.5.3 | MDA metagenomes are non-trivial | 82 |
| 3.5.4 | Metabolic co-dependence in the CCAP 1660/12 system | 83 |
| 3.6 | Conclusions | 84 |
| 4 | TRANSCRIPTOMIC ANALYSIS OF THE <i>PARAMECIUM BURSARIA</i> AND <i>MICRACTINIUM REISSESSI</i> ENDOSYMBIOSIS | 85 |
| 4.1 | Introduction | 85 |
| 4.2 | Aims | 88 |
| 4.3 | Methods | 88 |
| 4.3.1 | Sample preparation and sequencing | 88 |
| 4.3.1.1 | Bulk transcriptome RNA preparation | 88 |
| 4.3.1.2 | Single cell RNA preparation | 89 |
| 4.3.1.3 | Library preparation | 90 |
| 4.3.1.4 | Sequencing | 90 |
| 4.3.2 | Library contamination screening | 91 |
| 4.3.2.1 | Taxonomic analysis | 91 |
| 4.3.2.2 | GC density estimates | 92 |
| 4.3.3 | Optimising read pre-processing | 92 |
| 4.3.3.1 | Trimming | 92 |
| 4.3.3.2 | GC partitioning of reads | 93 |
| 4.3.3.3 | Error correction | 93 |
| 4.3.3.4 | K-mer normalisation and trimming | 93 |
| 4.3.4 | Assembly | 94 |
| 4.3.4.1 | Assembly assessment | 95 |
| 4.3.4.2 | ORF calling | 95 |
| 4.3.5 | Transcript binning | 95 |
| 4.3.5.1 | Initial BLAST based bins | 95 |
| 4.3.5.2 | Automated Phylogeny Generation Pipeline - Dendrogenous | 96 |
| 4.3.5.3 | Automated Phylogenetic Transcript Binning - Arboretum | 98 |
| 4.3.5.4 | TAXAssign comparison | 99 |
| 4.4 | Results | 99 |
| 4.4.1 | Library contamination screening | 99 |
| 4.4.2 | Read pre-processing | 105 |
| 4.4.2.1 | Trimming optimisation | 105 |
| 4.4.2.2 | GC partitioning | 107 |
| 4.4.2.3 | Error correction | 107 |
| 4.4.2.4 | Digital Normalisation | 109 |
| 4.4.3 | Assembly | 110 |
| 4.4.3.1 | Referenced | 110 |
| 4.4.3.2 | <i>de novo</i> assembly | 110 |
| 4.4.3.3 | Assembly combination | 111 |
| 4.4.4 | Binning | 112 |
| 4.4.4.1 | ORF calling | 112 |
| 4.4.4.2 | Performance of BLAST-based binning | 112 |
| 4.4.4.3 | Phylogeny-based bin classification | 112 |
| 4.4.4.4 | Performance relative to TAXAssign | 114 |

| | | |
|----------|--|------------|
| 4.5 | Discussion | 115 |
| 4.5.1 | Library screening is a key stage in sc-RNAseq | 115 |
| 4.5.2 | Combining single cell and bulk transcriptome data creates new challenges | 118 |
| 4.5.3 | Pre-assembly read partitioning is non-trivial | 119 |
| 4.5.4 | Digital normalisation greatly improves assemblies | 120 |
| 4.5.5 | Assembly and assembly assessment | 121 |
| 4.5.6 | Binning | 122 |
| 4.6 | Conclusion | 123 |
| 5 | METABOLIC INTEGRATION | 124 |
| 5.1 | Introduction | 124 |
| 5.1.1 | Metabolism of host and endosymbiont | 125 |
| 5.1.2 | Identifying direct points of contact | 127 |
| 5.1.2.1 | Transporter proteins | 127 |
| 5.1.2.2 | Secreted proteins | 128 |
| 5.1.3 | Metabolic mapping | 129 |
| 5.2 | Aims | 129 |
| 5.3 | Methods | 130 |
| 5.3.1 | Transporter analysis | 130 |
| 5.3.1.1 | Transporter identification pipeline | 130 |
| 5.3.2 | Secretome prediction | 130 |
| 5.3.3 | Qualitative expression analysis | 131 |
| 5.3.4 | Metabolic mapping analysis | 131 |
| 5.3.4.1 | <i>Chlorella variabilis</i> 1 N assembly | 132 |
| 5.3.5 | Metabolomics | 133 |
| 5.3.5.1 | Untargeted LC-QTOF profiling | 133 |
| 5.3.5.2 | Untargeted GC-QTOF profiling | 134 |
| 5.3.5.3 | Targeted amino acids quantitative analysis | 134 |
| 5.4 | Results | 135 |
| 5.4.1 | <i>P. bursaria</i> - <i>C. variabilis</i> Yad1g1N assembly | 135 |
| 5.4.2 | Transporter identification | 136 |
| 5.4.3 | Subset of transporters expressed in day and night | 137 |
| 5.4.4 | Secreted proteins | 137 |
| 5.4.5 | Metabolic maps | 139 |
| 5.4.6 | Metabolomics | 139 |
| 5.4.6.1 | Global profiling | 139 |
| 5.4.6.2 | Targeted amino acid analysis | 141 |
| 5.5 | Discussion | 146 |
| 5.5.1 | Novel sugars implicated in the endosymbiosis | 146 |
| 5.5.2 | Alternative exchanged amino acids in <i>P. bursaria</i> - <i>M. reisseri</i> | 147 |
| 5.5.3 | Potentially missing transporters and secreted proteins | 147 |
| 5.5.4 | Metabolomics shows promise | 149 |
| 5.6 | Conclusion | 150 |
| 6 | RNAi in <i>P. BURSARIA</i> | 151 |
| 6.1 | Introduction | 151 |
| 6.1.1 | RNAi | 151 |
| 6.1.2 | RNAi in <i>Paramecium</i> | 153 |
| 6.1.3 | RNAi “cross-talk” | 154 |
| 6.2 | Aims | 155 |
| 6.3 | Methods | 155 |
| 6.3.1 | RNAi constructs | 155 |
| 6.3.1.1 | RNAi feeding | 155 |
| 6.3.1.2 | RNAi microinjection | 156 |
| 6.3.2 | Analysis of RNAi pathway | 156 |
| 6.3.2.1 | Survey for RNAi components in <i>P. bursaria</i> | 156 |
| 6.3.2.2 | Phylogenetic analysis of RNAi pathway | 156 |
| 6.3.2.3 | Structural prediction and functional analysis | 157 |
| 6.3.3 | dsRNA cross-talk analysis - “eDicer” | 157 |
| 6.4 | Results | 158 |
| 6.4.1 | Induction of RNAi | 158 |

| | | |
|---------|--|------------|
| 6.4.2 | RNAi pathway components | 159 |
| 6.4.2.1 | Dcr1 | 159 |
| 6.4.2.2 | Pds1 | 159 |
| 6.4.2.3 | Cid | 159 |
| 6.4.2.4 | Rdr | 162 |
| 6.4.2.5 | Piwi | 162 |
| 6.4.3 | dsRNA collisions | 162 |
| 6.5 | Discussion | 169 |
| 6.5.1 | No dsRNA RNAi inducible phenotypes in <i>P. bursaria</i> SW1 | 169 |
| 6.5.2 | Missing components of RNAi pathway in <i>P. bursaria</i> | 170 |
| 6.5.3 | Endosymbiont “collision” hypothesis | 172 |
| 6.6 | Conclusions | 173 |
| 7 | CONCLUSIONS AND RECOMMENDATIONS | 174 |
| | APPENDICES | 224 |
| A | APPENDIX 1 | 225 |
| A.1 | Arboretum classifier comparison | 225 |
| A.1.1 | Genomes used | 225 |
| A.1.2 | Classification reports | 227 |
| A.2 | eDicer | 229 |
| A.2.1 | Eukaryote transcriptomes | 229 |
| A.2.2 | Bacterial transcriptomes | 231 |
| A.2.3 | Archaea transcriptomes | 236 |

List of Figures

| | |
|--|-----|
| 1.1.1 The spectrum of endosymbioses | 17 |
| 1.2.1 Christiaan Huygens: the discoverer of <i>P. bursaria</i> | 20 |
| 1.2.2 <i>Paramecium</i> Genomes | 21 |
| 1.2.3 Life cycle of <i>P. bursaria</i> | 23 |
| 1.2.4 Taxonomic context of host and endosymbiont | 25 |
| 1.3.1 Established of endosymbiosis in <i>P. bursaria</i> | 30 |
| 2.2.1 Illumina library preparation | 35 |
| 2.2.2 Overview of paired-end sequencing | 36 |
| 2.2.3 Compounds by MS Separation Method | 42 |
| 2.2.4 Mass Spectrometry Data | 43 |
| 2.3.1 Explanation of learning curves and fitting | 45 |
| 2.3.2 SVM decision boundaries | 47 |
| 2.3.3 Soft-margin classifiers | 48 |
| 2.3.4 Kernel trick | 49 |
| 2.3.5 K-means clustering | 51 |
| 3.1.1 ITS ₂ structure overview | 65 |
| 3.3.1 ITS ₂ primer locations | 67 |
| 3.4.1 ITS ₂ phylogeny | 72 |
| 3.4.2 ITS ₂ SNP alignment | 73 |
| 3.4.3 Genome assembly GC densities | 74 |
| 3.4.4 Genome Assembly Cumulative Contig Lengths | 75 |
| 3.4.5 Genome assembly gaps | 76 |
| 3.4.6 Genomic contig clustering | 78 |
| 3.4.7 Endosymbiont variant count and frequency density | 80 |
| 4.3.1 Tree generation pipeline architectures | 97 |
| 4.4.1 GC densities of libraries | 100 |
| 4.4.2 GC densities of <i>P. bursaria</i> Yad1g - <i>C. variabilis</i> 1N | 101 |
| 4.4.3 GC Densities of Bulk Libraries | 101 |
| 4.4.4 Visualisation of taxonomically excluded libraries | 103 |
| 4.4.5 Visualisation of taxonomically included libraries | 103 |
| 4.4.6 Comparison of Subsample to Whole Library Profiles | 104 |
| 4.4.7 Optimising trimming thresholds | 106 |
| 4.4.8 Plot of parKour clusters | 108 |
| 4.4.9 Preliminary transcript binning analysis | 113 |
| 4.4.10 Radial visualisation of training data | 114 |
| 4.4.11 Radial Visualisation of Test and Training Data | 115 |
| 4.4.12 F1-scores of different classifiers | 116 |
| 4.4.13 Normalised confusion matrix of transcript classification | 116 |
| 5.4.1 Summary of <i>M. reisseri</i> endosymbiont bin annotation | 136 |
| 5.4.2 Jasper-shannon divergence of single cell transcriptome libraries | 137 |
| 5.4.3 A binary filter heatmap that displays the 4 groups of transporter proteins identified in the <i>P. bursaria</i> - <i>M. reisseri</i> transcriptome. Specifically, it shows the way that transporters fall into different expression “groups” and their relative sizes. In detail, 34 expressed only in dark single cell libraries, 48 expressed only in light libraries, the 64 expressed in both and the 14 only recovered in the bulk libraries. | 139 |
| 5.4.4 KEGG maps of endosymbiont bin compared with other algae | 140 |
| 5.4.5 GC-QTOF cloud plot | 141 |

| | |
|---|-----|
| 5.4.6 Plot of identifiable GC-QTOF metabolites | 142 |
| 5.4.7 LC-QTOF cloud plots | 143 |
| 5.4.8 Of the 254 positive significantly different present metabolites, 19 were removed after manual inspection of peaks, 95 were removed due to having no METLIN hits | 144 |
| 5.4.9 254 positive significantly different present metabolites, 19 were removed after manual inspection of peaks, 95 were removed due to having no METLIN hits | 144 |
| 5.4.10 43 significantly different present metabolites were identified, 3 were removed after manual inspection of peaks, 17 were removed due to having no METLIN hits | 144 |
| 5.4.11 LC-QQQ Quantitative Analysis of Amino Acids | 145 |
| | |
| 6.4.1 Dcr1 Phylogeny | 160 |
| 6.4.2 Pds1 Phylogeny | 161 |
| 6.4.3 Predicted Structure of Pds1 | 161 |
| 6.4.4 Cid Phylogeny | 163 |
| 6.4.5 Rdr1 and Rdr2 Phylogeny | 164 |
| 6.4.6 Rdr3 Phylogeny | 165 |
| 6.4.7 Unique eDicer Collisions by Subject Class | 166 |
| 6.4.8 Pair-plot of Normalised Collisions in Both Endosymbiont | 167 |
| 6.4.9 Pair-plot of Collisions in Host-Endosymbiont pairs | 167 |
| 6.4.10 Regression of eDicer Collisions by Query Length | 168 |
| 6.4.11 Normalised Unique eDicer Collisions | 169 |
| 6.5.1 DAPI Stained <i>P. bursaria</i> MAC | 170 |
| 6.5.2 Summary of RNAi Factors in <i>Paramecium</i> species | 171 |
| 6.5.3 Summary of RNAi Factors Evolutionary Scenarios in <i>Paramecium</i> species | 172 |

List of Tables

| | |
|--|-----|
| 1.1.1 Types of Biological Interaction | 16 |
| 3.4.1 The effect of read trimming threshold across libraries | 73 |
| 3.4.2 Genome Assembly Statistics | 74 |
| 3.4.3 Taxonomic Assignment of Genomic Contigs | 77 |
| 3.4.4 Explanation of Clustering Errors | 77 |
| 3.4.5 Custom Taxonomic Binning | 79 |
| 4.4.1 DueyDrop Taxonomic Profile Summary | 102 |
| 4.4.2 Taxonomic Profiles of Bulk Libraries | 102 |
| 4.4.3 Comparison of Effect of Trimming on Trinity Assemblies | 105 |
| 4.4.4 Library Sizes After Trimming | 105 |
| 4.4.5 ParKour Cluster Summaries | 107 |
| 4.4.6 Effect of Error Correction on Assembly | 108 |
| 4.4.7 Effect of Digital Normalisation on Assembly | 109 |
| 4.4.8 Summary of Different Assembler Outputs | 110 |
| 4.4.9 Trinity vs. Bridger Assemblies | 111 |
| 4.4.10 Merged Assembly Summary | 111 |
| 4.4.11 Classification Report for K-Neighbours | 114 |
| 4.4.12 TAXAssign Taxonomic Assignments | 117 |
| 4.6.1 Summary of Transcriptome Bins | 123 |
| 5.4.1 Results of various types of pre-processing for the <i>Yad1g1N</i> transcriptome | 135 |
| 5.4.2 Summary of <i>Yad1g1N</i> Transcriptome Assemblies | 135 |
| 5.4.3 <i>Yad1g1N</i> Transcript Binning Summary | 135 |
| 5.4.4 Summary of Predicted Transporters Across Algal Sequences | 136 |
| 5.4.5 List of Transporters Present in Both Light and Dark SCT Libraries | 138 |
| 6.1.1 Summary of RNAi pathway components in <i>P. tetraurelia</i> | 153 |
| 6.3.1 Table of RNAi vectors and inserts used | 155 |
| 6.3.2 RNAi pathway components from (Marker et al., 2014) | 157 |
| A.1.1 KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_neighbors=50, p=2, weights='uniform') | 227 |
| A.1.2 LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.0001, verbose=0) | 227 |
| A.1.3 SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, degree=3, gamma=0.0, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False) | 227 |
| A.1.4 DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, random_state=None, splitter='best') | 227 |
| A.1.5 ExtraTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, random_state=None, splitter='random') | 228 |
| A.1.6 RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False) | 228 |
| A.1.7 AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None) | 228 |
| A.1.8 LDA(n_components=None, priors=None, shrinkage=None, solver='svd', store_covariance=False, tol=0.0001) | 228 |
| A.1.9 QDA(priors=None, reg_param=0.0) | 228 |

| | |
|--|-----|
| A.1.1 GaussianNB() | 228 |
| A.1.1 LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0) | 229 |

Acknowledgements

I'd like to thank Thomas Richards for his extensive guidance and support during this PhD. Not many supervisors would allow such freedom to pursue diverse avenues of learning and research even when those avenues might have been at odds with the research in question.

Secondly, I'd like to thank Gavin Gray for his constant end-to-end encrypted presence, provision of chip-tunes, proof-reading services, and distraction with machine learning side-projects.

Coffee, and its excellent providers at "Camper Coffee" has played a significant role in the completion of this thesis. I'd also like to thank Aurelie Chambouvet and Adam Monier for their frequent company on coffee forays.

Sarah Gurr for both accepting my application and encouraging my interests throughout my undergraduate degree and as the puppetmaster that masterminded my eventual post-graduate studies.

Guy Leonard for advice and guidance in my first forays into bioinformatics. David Milner for his considerable contributions to the lab work in this thesis. Theresa Hudson and Ines Yang for help in the culturing and maintenance of *Paramecium bursaria* cultures.

Finally, but by no means least, thanks to my parents for their sustained support throughout my academic career and life so far.

Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's law

- Douglas Hofstadter: *Gödel, Escher, Bach: An Eternal Golden*

Braid, 1979

1

Introduction

1.1 ENDOSYMBIOSIS

1.1.1 WHAT IS ENDOSYMBIOSIS?

Endosymbiosis has proven one of the most fundamental processes in the evolution of the eukaryotic cell (Timmis et al., 2004; Lane, 2007; Martin and Herrmann, 1998; Archibald, 2015). It has both shaped the global climate and created the cellular context in which specialised multicellular organisms have evolved.

Endosymbiosis is a special case of symbiosis, which results in a long-term stable interdependent living together (“sym/σύν” – together, “bios/βίωσις” – living) of two or more organisms to a point of mutual benefit (de Bary, 1869; Pound, 1893) (although many now expand this definition beyond mutualism to include other categories of biological interactions (Leung and Poulin, 2008; O’Malley, 2015)). What differentiates endosymbiosis from symbiosis in general is that one partner (the endosymbiont) lives wholly inside (“endo/ἔνδον” - inside) of another (the host). This “inside” can refer to symbionts either living intracellularly or within the tissues of multicellular organisms. However, it excludes niches such as the digestive tract of metazoa as this can be considered as an external surface of the host. These latter symbionts are occasionally termed ectosymbionts.

There is a considerable diversity of endosymbiotic relationships in nature. These relationships can encompass many different degrees of host-symbiont integration, interdependence and ecological interaction types (see

| Interaction Name | Interaction Outcome |
|------------------|---------------------|
| Mutualism | (+, +) |
| Antagonism | (+, -) |
| Competition | (-, -) |
| Commensalism | (+, o) |
| Amensalism | (-, o) |
| Neutralism | (o, o) |

Table 1.1.1: An overview of the categories of biological interaction and the effect they have on the two interacting biological units, which may be anything from individual species to whole populations. The outcome column contains a tuple relating the effect an interaction has on a pair of interacting biological units. This “effect” is often assessed in terms of metrics such as individual fitness, population size and/or growth rate. Note: parasitism and predation are mechanisms by which an antagonistic interaction may take place (Abrams, 1987) in the same sense that endosymbiosis is a mechanism by which a mutualistic interaction can take place. In reality most interactions will not fall neatly into one of these categories and throughout its duration will often display characteristics of multiple categories (Leung and Poulin, 2008)

table 1.1.1). Even if we restrict ourselves to endosymbioses that are largely “mutualistic” (noting that the exact nature of a certain endosymbiosis is highly dependent on the specific ecological context at a particular point of time and doesn’t always neatly quantise into discrete categories (Leung and Poulin, 2008)) there is a broad range of characteristics.

It is worth briefly addressing a common motif of biology: the application of discrete schemas to continuous distributions. These biological quantisations are prone to error (fuzzy delineations) and are constantly challenged by novel discoveries which exhibit a mosaic of category features. There are many examples of this such as the classification of mitochondria-related organelles (Maguire and Richards, 2014), types of biological interactions (see table 1.1.1), and the numerous species concepts (De Queiroz, 2007; Boenigk et al., 2012). That is not to say biological quantisation is without utility or is a futile task. Indeed, as long as there is a clarity to the application, basis and limitations of these schema then they form a critical (epistemological) framework upon which further research and communication can build (Boenigk et al., 2012). However, care must be taken not to forget that they do not reflect reality and can inadvertently obscure the grey areas (Leung and Poulin, 2008).

For example, in terms of interdependence of host and endosymbiont you could construct a spectrum with “incidental” endosymbioses such as bacterial or fungal escape of digestion in macrophages at one extreme and at the other obligate systems such as the mitochondria or chloroplast where host and symbiont are essentially a single organism (Margulis, 1998; Archibald, 2015). In the middle of such a spectrum you could find facultative endosymbioses where each partner is capable of, and does, live aposymbiotically for extended life phases e.g. Rhizobia soil bacteria and legume (Fabaceae) plants (reduction of atmospheric N₂ to ammonia (Hirsch, 1992) in exchange for host-derived carbon sources such as malate and succinate (Prell and Poole, 2006)).

An endosymbiosis may be highly integrated in terms of metabolism, genome and life history while still only being moderately interdependent (such as the facultative Rhizobia nitrogen fixation which takes place in carefully controlled specialised root nodule structures (Crespi and Frugier, 2008)). However, generally interdependence and integration correlate reasonably well due to the increased selective pressure to minimise lethal aberrant interactions that comes with interdependence. This can be seen in the extreme of host-symbiont integration: that of the endosymbiotic organelles, which are so highly integrated they were considered part of the cell by mainstream

scientific establishment until only relatively recently (Archibald, 2014).

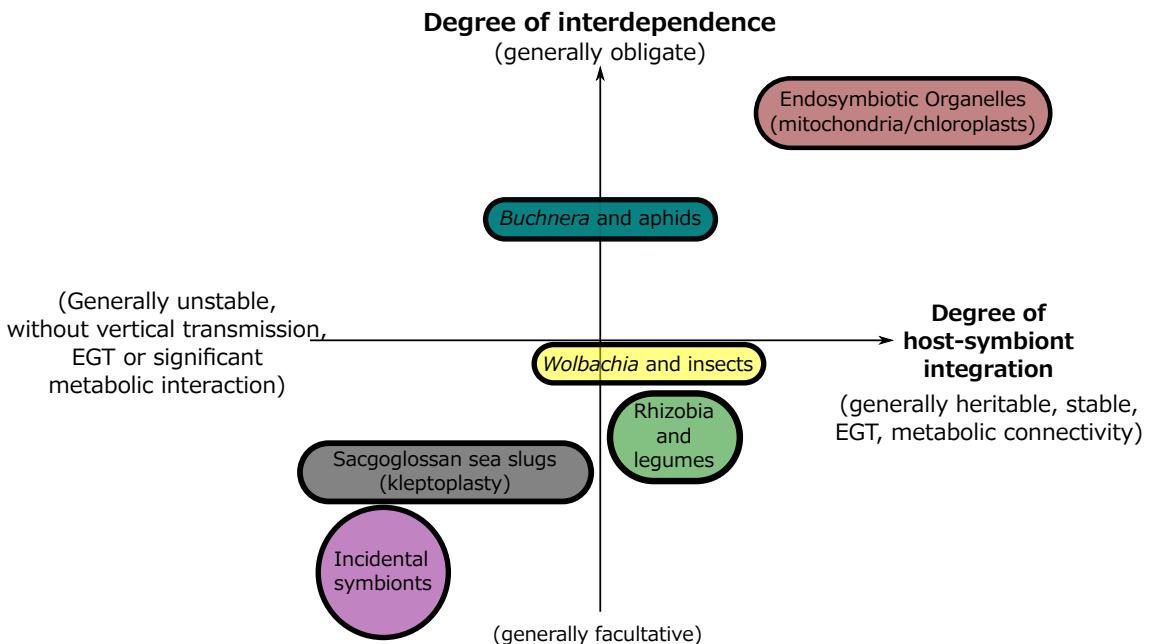


Figure 1.1.1: Plot demonstrating a fragment of the diversity of endosymbioses and specifically highlighting the possibility of a well integrated by facultative endosymbiosis. Host-symbiont integration is a rough measure of how connected the host and symbiont have become genomically, metabolically and in terms of life history. Whereas, interdependence is a approximate measure of the degree to which the relationship is necessary for life of organisms involved. It should be noted that both axes can be highly reliant on specific ecological and environmental context.

Intracellular endosymbionts can be found inhabiting multiple host-compartments from nakedly in the cytoplasm, to host-derived vacuolar compartments (often from exo- and endocytic systems) (e.g. (Kodama and Fujishima, 2009)) along with a range of host organelles including the endoplasmic reticulum (Vogt, 1992), Golgi body (Cho et al., 2011), mitochondria Sasser et al. (2006), chloroplast (Wilcox, 1986) as well as the nucleus (first discovered in *Paramecium* (Schulz and Horn, 2015)). Owing to the endosymbiotic origin of the chloroplast and mitochondria it becomes apparent that there can be multiple “layers” of endosymbiosis. A primary endosymbiont is an endosymbiont that is the direct endosymbiont of the host (e.g. the mitochondria to eukaryotes) whereas a secondary endosymbiont is the endosymbiont of an endosymbiont. The layers of endosymbioses can get impressively deep, for example, bacterial endosymbionts have been identified within the chloroplast stroma (cyanobacterial endosymbiont) of dinoflagellates (e.g. *Woloszynskia pascheri* (Wilcox, 1986)) In turn, dinoflagellate plastids have been discovered that are likely the product of tertiary endosymbioses (Gabrielsen et al., 2011) with higher-order events hypothesised in related groups (Stiller et al., 2014). Therefore, bacteria like this could be the endosymbiont of an endosymbiont of an endosymbiont of an endosymbiont (quaternary) or higher.

With this considerable diversity it is perhaps not surprising that endosymbioses have been discovered featuring partners from all 3 domains of cellular life. However, with the exception of one extant Bacteria-Bacteria endosymbiosis (von Dohlen et al., 2001), typically the majority of known endosymbioses feature a eukaryotic

host¹ but can include endosymbionts from all 3 domains. For example:²

- Eukaryote-Archaea ([Moissl-Eichinger and Huber, 2011](#))
 - Methanogenic archaea within various ciliates species (e.g. *Plagioplyxa frontata*) ([Fenchel and Finlay, 1992](#); [Lange et al., 2005](#))
 - *Cenarchaeum symbiosum* within the tissues of marine sponges ([Preston et al., 1996](#); [Wrede et al., 2012](#))
- Eukaryote-Bacteria
 - *Hartmannella* and it's intranuclear endosymbiont *Candidatus Nucleicultrix amoebiphilia* ([Schulz et al., 2014](#))
 - the most famous pairing of mitochondria and plastids
- Eukaryote-Eukaryote
 - The fungi *Diplodia mutila* which aids herbivory resistance in the palm *Iriartea deltoidea* in lowlight conditions but becomes pathogenic if host is well lit ([Álvarez Loayza et al., 2011](#))
 - Red alga derived plastids in brown algae ([Dorrell and Smith, 2011](#))
 - Numerous examples of algal mediated acquired phototrophy in ciliates ([Johnson, 2011](#))

Endosymbiosis is the one of the most significant evolutionary processes in eukaryotic cell. It offers a means for eukaryotes to benefit from the extensive metabolic diversity present in the bacterial and archael pangenome, especially the only known forms of primary energy production - photosynthesis and chemosynthesis ([Werneck, 2012](#)).

1.1.2 PLASTID ENDOSYMBIOSES

Most molecular evidence currently points towards a single primary endosymbiotic event between a phagotrophic ancestral eukaryote (with mitochondria and developed endomembrane system ([Rockwell et al., 2014](#))) and a cyanobacteria (blue-green algae) as giving rise to the archaeoplastida (that is the green algae, red algae, glauco-phytes and land plants ([Green, 2011](#))) and their double membrane bound plastids ([Keeling, 2013](#)). While, this event is one of the must fundamental events in the evolution of life in and of itself it is only capable of explaining

¹There are however many examples of mutualistic symbioses which are Bacteria-Bacteria (e.g. biofilms ([Watnick and Kolter, 2000](#))), Bacteria-Archaea (e.g. anaerobic methanotrophic archaea and sulphate-reducing bacteria likely responsible for a large proportion of global methane consumption ([Boetius et al., 2000; Knittel and Boetius, 2009](#)) and SM1 euryarchaea/*Thiotricha* sp. sulphide-oxidising bacteria ([Henneberger et al., 2006; Wrede et al., 2012](#))), and at least one example of Archaea-Archaea (*Igniococcus hospitalis*/*Nanoarchaeum equitans* ([Huber et al., 2002](#)). Interestingly *Igniococcus* is the first identified case of an energised outer-membrane in double-membrane bound archaea or bacteria, a significant finding for the development of theories of eukaryogenesis ([Küper et al., 2010](#)))

²Although with all these example, it is important not to consider an endosymbiotic relationship in isolation from other endosymbionts present in the same host. There are examples where facultative "secondary endosymbionts" are able to compensate for the loss of an obligate endosymbiont ([Koga et al., 2003](#)). Symbiont-symbiont interactions have been found to play a role in determining which endosymbionts are capable of establishing themselves in a certain host and can even be capable of generating additional phenotypes e.g. the R-bodies of "killer" *Paramecium* species which may be a product of an interaction between the *Paramecium* host, *Caedibacter* and a bacteriophage ([Schrallhammer and Schweikert, 2009](#)).

a small proportion of the diversity of plastids across the eTOL (Keeling, 2013) Apart from one other putative primary endosymbiosis in *Paulinella chromatophora* (a euglyphid amoeba with photosynthetic chromatophores that are vertically inherited, synchronised to host and bear a much stronger molecular and morphological resemblance to reduced cyanobacteria than the chloroplast of the archaeplastida (Kies and Kremer, 1979; McFadden, 2014)) all other oxygenic phototrophs (as well as several non-photosynthetic but plastid bearing pathogens (Sato, 2011)) have arisen by secondary or higher order endosymbioses (Hoshina and Imamura, 2009). Secondary endosymbioses are those in which another eukaryotic lineage has engulfed a primary plastid bearing algae and reduced and integrated them in a simulacrum of primary endosymbiosis, occasionally serially (Keeling, 2010). This and subsequent loss of membranes leads to the range of membrane layer numbers around plastids in various eukaryote lineages (Keeling, 2013). This secondary order plastid endosymbioses have occurred independently in divergent eukaryote lineages e.g. chloroarachniophytes and euglenids, and an unresolved number of times in the set of cryptomonads, haptophytes, stramenopiles, dinoflagellates and apicomplexans (Keeling, 2013). As well as an uncertain number of higher order endosymbioses in the dinoflagellates (Keeling, 2013).

Therefore, understanding the mechanisms and evolution of secondary photosynthetic endosymbioses would provide important insight into the evolution of a considerable number of eukaryotic lineages. Unfortunately, most extant examples feature endosymbioses within which metabolic co-dependence has already become fixed masking the potential mechanisms through the endosymbiosis may have originated. Facultative systems such as the *Chlorella* endosymbionts of *Paramecium bursaria* offer a potential avenue to investigate secondary photosynthetic endosymbioses at an earlier stage before metabolic co-dependence has become fixed (while acknowledging the impossibility of interrogating events that have already occurred within the correct ecological context. Furthermore, as the ancestral protist involved in the primary plastid endosymbiosis likely exhibited a similar life style to serially phagotrophic *Paramecium* and would initially at least have been mixotrophic (combining phagotrophy with phototrophy via the newly acquired plastid (Rockwell et al., 2014) in the same manner as *Paramecium bursaria* (and other mixotrophic ciliates (Johnson, 2011)) the study of the *Paramecium bursaria-Chlorella* system offers potential insight into this early and fundamental stage of eukaryote evolution.

1.2 PARAMECIUM BURSARIA

Paramecium are large (50 – 330 μ m) phagotrophic single-celled eukaryotes belonging to a genetically diverse (Prescott, 1994) sub-grouping of the alveolates known as the ciliates (see 1.2.4). They have been studied since the invention of microscopy (Görtz and Fokin, 2009) (first recorded by a contemporary of van Leeuwenhoek see 1.2.1) and are some of the longest-standing model unicellular eukaryotes. They have been used to study everything from mutagenesis and developmental genetics, to genomics rearrangement and epigenetics (McGrath et al., 2014). As such have a well-developed methodological (Sonneborn, 1970) and theoretical literature along with several available genomes (see 1.2.2 for genomes and their relative relationship to *P. bursaria*).

Paramecium bursaria “the green *Paramecium*” is distinguished from most³other *Paramecium* by the distinc-

³There is at least one other species, *Paramecium chlorelligerum*, that harbours a different green algae (*Meyerella*) (Kreutz et al., 2012) and owing to the multiple origins of algal symbionts in *P. bursaria* (Hoshina and Imamura, 2009) and the general prevalence of mixotro-

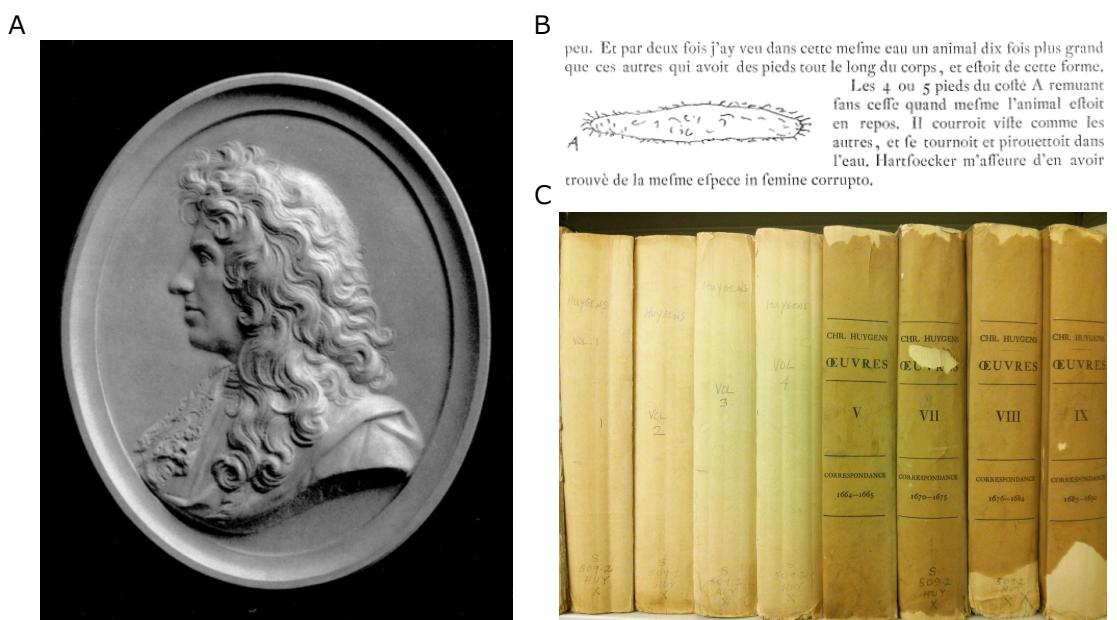


Figure 1.2.1: **A:** Carving of Christiaan Huygens (1629-1695), the prominent Dutch Golden Age mathematician and scientist and contemporary of Antoni van Leeuwenhoek, from a medallion by Jean-Jacques Cléron 1679 (reproduced from (Huygens, 1899)). **B:** Likely the first sketch of the micro-organism that we now know as *Paramecium* by Christiaan Huygens in a letter (No. 2133, 11th of August 1678) to his father Constantijn Huygens. An approximate translation of the accompanying text goes as follows “I have twice seen in this water an animal 10 times as large as the others and with feet all over its body and a narrow form. 4 or 5 feet stirred even when the animal was at rest. It moves as fast as the others, turning and spinning in the water. Hartfoecker thinks he may have discovered the same species in ‘semine corrupto’ (as a dried out husk?).” (reproduced from (Huygens, 1899)). **C:** 8 of the 10 volumes of the collected correspondences of Christian Huygens as prepared for the Dutch Society of Sciences and published from 1888-1905.

tive stable, heritable secondary photosynthetic endosymbiosis it maintains with several species of the green algae *Chlorella*. Each *P. bursaria* 100 – 160 μm (Jennings, 1939) cell contains ~ 300 endosymbiotic algae maintained in individual perialgal vacuoles (PV) around the cell cortex (Hoshina and Imamura, 2009).

Much like other ciliates, *Paramecium* are covered by cilia. These are minute hairlike biochemically heterogeneous organelles capable of sensing the environment and by beating in co-ordinated metachronal waves of power strokes and recovery strokes (Funfak et al., 2015) provide cellular locomotion and, in the case of phagotrophic ciliates like *Paramecium*, forcing food bacteria towards the oral groove (cytopharynx) where they can be phagocytosed (Hamel et al., 2011; Aubusson-Fleury et al., 2015). *Paramecium* is also capable of rapid locomotion via the expulsion of trichocysts. These are defensive membrane-bound organelles known as trichocysts containing a crystalline spike which can be rapidly ejected into the environment on fusion of trichocyst membrane with plasma membrane (Hamel et al., 2011).

Like other ciliates, including *Tetrahymena*, *Paramecium* deviates from the universal genetic code. Canonical stop codons TAA and TAG have been reassigned to produce glutamine therefore there is only one stop codon (TGA) but 4 glutamine codons (Salim et al., 2008).

Another defining feature of *Paramecium*, and ciliates in general, is a unique means of germline sequestration from somatic function in the form of “nuclear dimorphism” (Jahn and Klobutcher, 2002). Specifically, they have phy in ciliates (Johnson, 2011) there are likely others yet to be discovered.

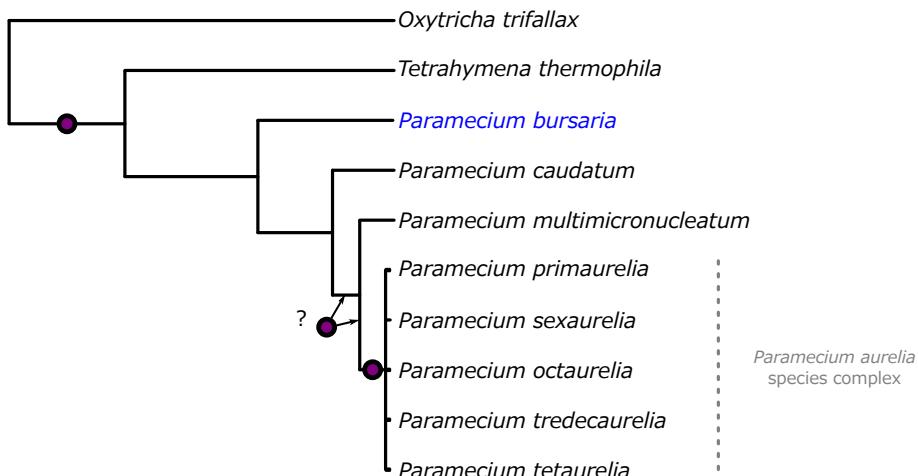


Figure 1.2.2: Modified phylogeny redrawn from (Fokin et al., 2004; Aury et al., 2006; McGrath et al., 2014) Showing the relative relations of ciliate species with genomic/transcriptomic resources and hypothesised WGD event locations with a purple dot. Specifically 2 strains of *Paramecium tetraurelia* (Aury et al., 2006), assemblies for *P. caudatum*, *P. multimicronucleatum*, and *P. aurelia* complex species *P. sexaurelia*, *P. primaurelia*, *P. octaurelia* and *P. tredecaurelia* on ParameciumDB (as of 05/03/2015) (Arnaiz and Sperling, 2011a). As well as *Tetrahymena thermophila* (Eisen et al., 2006) and *Oxytricha trifallax* (Swart et al., 2013).

two types of nuclei, expression optimised highly polyploid somatic macronuclei (MAC) and largely silent diploid germline micronuclei (MIC) ([Prescott, 1994](#)).

During normal vegetative growth the MIC is densely packed with chromatin, is transcriptionally silent and undergoes mitosis as standard. Meanwhile, the MAC reproduces by a non-standard pinching process termed “amitosis”. This process appears to lack any mechanism to ensure equal segregation of chromosomes such as spindle fibres (Chalker et al., 2013). On the other hand, during sexual reproduction (in which two compatible *P. bursaria* exchange haploid MIC gametes generated by meiosis) the MAC degrades and must be reconstituted entirely from the newly formed heterozygous MIC (Jahn and Klobutcher, 2002) (see 1.2.3). The exact number of MIC and MAC varies widely by species and genus however, *P. bursaria* contains a single large MIC which consists of 80 to several hundred chromosomes depending on the exact subspecies (Chen, 1940).

P. bursaria reaches sexual maturity after 50-100 fissions (Siegel and Larison, 1960) and will conjugate with another *Paramecium bursaria* cell of compatible mating type and exchange haploid MIC gametes. Most *Paramecium* have a finite number of vegetative divisions and will die if they do not sexually reproduce (Chalker et al., 2013). Unlike all other studied *Paramecium*, *P. bursaria* does not only have 2 mating types and appears to have undergone gene duplication at two unlinked mating type loci. Different *P. bursaria* isolates display 2, 4 and 8 mating types (Phadke and Zufall, 2009) and form 4 or more mutually incompatible groups (Jennings, 1939). Sexually *P. bursaria* appears to have synclonal inheritance - a strictly mendelian inheritance contrary to the more complex epigenetic patterns observed in other *Paramecium* species (that led to much of the early work on epigenetics) (Siegel and Larison, 1960; Phadke and Zufall, 2009). During conjugation there is minimal cytoplasmic exchange (with no exchange of endosymbionts) (Wichterman, 1946). Contrary to other *Paramecium* species, which undergo autogamy after 75 rapid replications (Sung et al., 2012) or 30-35 while starving (Berger, 1986), *Paramecium bursaria* has not been found to naturally undergo autogamy (Siegel, 1963; Yanagi, 2004) but it can be induced by treatment of methyl cellulose (Yanagi, 2004)

In *Paramecium* the haploid size and complexity of the MIC is greater than that of the MAC as a result elimination of approximately 20-30% of DNA during reconstitution of the MAC from the MIC. Eliminated sequences are known as internal eliminated sequences (IES) and are a mixture of transposon-related repetitive sequences and nongenic single-copy sequences resulting in a gene dense low-repeat MAC (Chalker et al., 2013). Similarly, the MAC has a greater number of shorter chromosomes than the MIC due to chromosomal fragmentation during DNA elimination by imprecise deletion of internal DNA segments followed by rejoicing or telomere addition (Chalker et al., 2013). This process involves 3 steps is observed in *P. tetaurelia* and features a special class of small RNAs (scnRNAs) (Chalker et al., 2013):

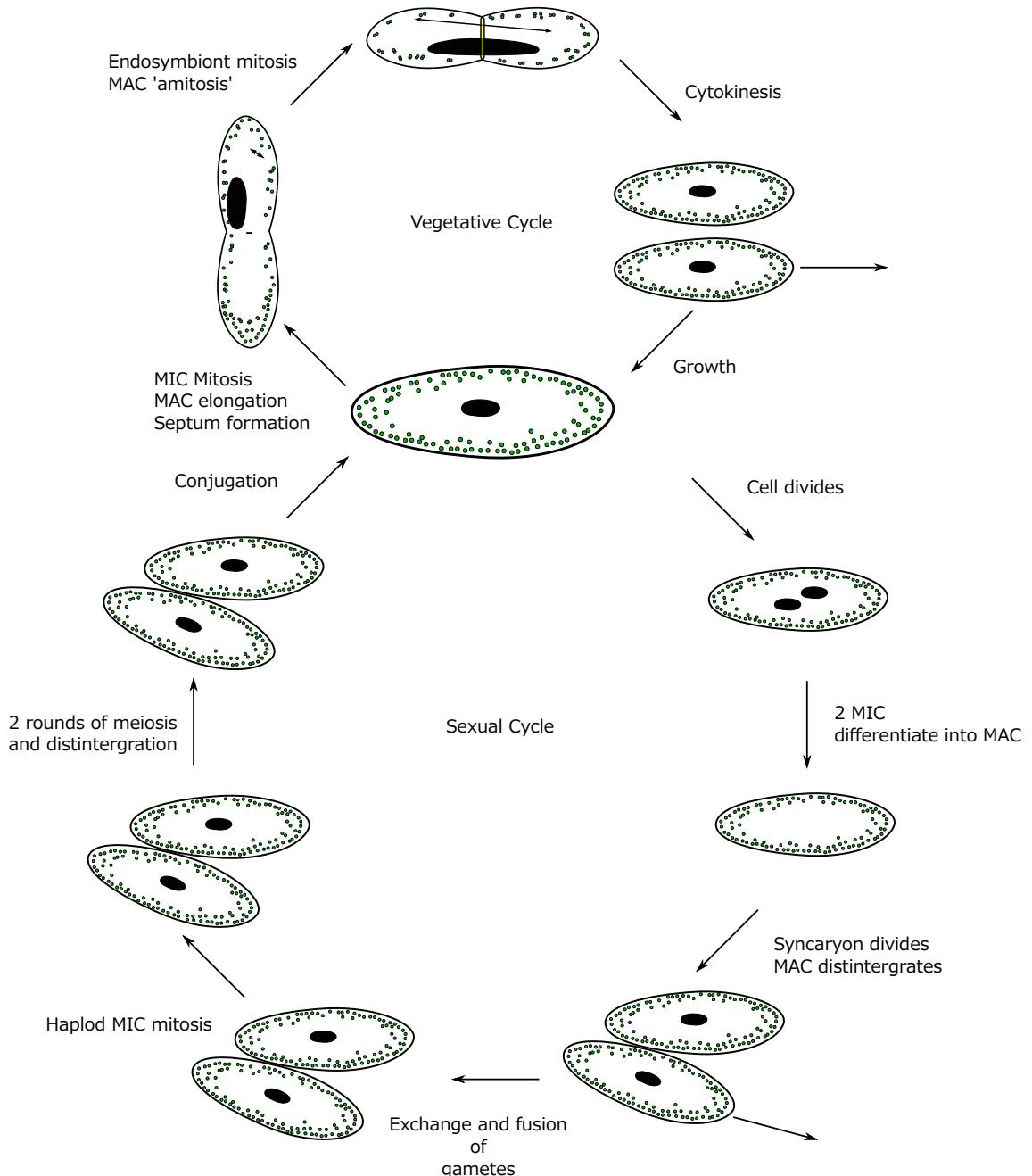
- DNA amplification to high ploidy.
- DNA elimination pathway 1: accurate removal of short unique-copy elements (IES, internal eliminated sequences) that run through coding and non-coding sequences. This is achieved using bounding 5'-TA-3' dinucleotides to target double-stranded breaks and subsequent end-joining (Mayer and Forney, 1999; Bétermier, 2004)
- DNA elimination pathway 2: imprecise removal of large DNA regions (often containing transposable elements) in a manner similar to transposon silencing in other eukaryotes. This process likely involves short ncRNAs targeting heterochromatin formation via histone methylation to induce fragmentation. This fragmentation is subsequently repaired via the addition of new telomeres (Duret et al., 2008).

This process involves a special class of meiosis specific small RNAs (scnRNAs) which target aspects of DNA elimination (Chalker et al., 2013).

While the MIC appears to vary in size between subspecies of *P. bursaria* the MAC is roughly the same size and generally contains 10-30 times the amount of DNA than the diploid MIC (Cullis, 1972). Therefore, the MAC of *P. bursaria* is likely 20-60n (in contrast to the 80n MAC ploidy found in *P. tetaurelia* (Duret et al., 2008)). The MAC genome is likely to be somewhere between 20 and 100Mb and contain somewhere between 18,000 and 40,000 genes based on the size of the *Tetrahymena* (Eisen et al., 2006), *P. tetaurelia* (Aury et al., 2006), *P. caudatum* (McGrath et al., 2014) and *Oxytricha* (Swart et al., 2013) MAC genomes.

We can infer other likely features of the *P. bursaria* MAC genome from the *P. tetaurelia* sequencing project. Specifically, it is likely AT-rich (28% GC in *P. tetaurelia*), compact (78% coding density in *P. tetaurelia*), mostly repeat free with small intergenic regions and many short introns (e.g. 25bp IES elements) (Aury et al., 2006). There is also likely to be evidence of at least 1 whole genome duplications (WGD) (an ancient WGD before the divergence of *Tetrahymena* and *Paramecium* clades but not the 2 most recent WGD (see fig. 1.2.2) giving rise to the *P. aurelia* complex) with a moderate level of conservation to gene synteny and duplicated gene retention (weakly correlating with a genes GC%, expression level and functional class) (Aury et al., 2006; McGrath et al., 2014). *P. bursaria* is also likely to have a high level of replication fidelity and relatively low rate of base-substitution mutation, traits found in *P. tetaurelia*, as *P. bursaria* shares ciliate specific modifications to the active sites of B-family polymerases α , ζ , and the proofreading exonuclease of DNA polymerase ε believed to be adaptions to improve

Figure 1.2.3: Figure redrawn and modified from (Duret et al., 2008). During normal vegetative growth *Paramecium bursaria* (and other *Paramecia*) divide by binary fission with the MAC elongating and “pinching” off in a process distinct from mitosis (known as amitosis) while the MIC undergoes mitosis. As the cell pinches before cytokinesis an unknown septum forms at the “neck”, this stops cytoplasmic streaming which induces the endosymbionts to begin to divide. Cytokinesis then occurs largely simultaneously in host and endosymbionts (Kadono et al., 2004; Takahashi et al., 2007). The sexual cycle involves conjugation of compatible mating types (taking around an hour and lasts 24-48 hours (Jennings, 1939)) which triggers two-rounds of meiosis of the MIC with one product disintegrating after each division so only a single haploid MIC remains. This undergoes mitosis to produce male and female gametes. Male gametes are then reciprocally exchanged between mating cells and fuse with the respective female gamete to create a syncaryon. Each syncaryon divides once and one product disintegrates before undergoing two subsequent divisions. Two products differentiate into MACs by programmatic reorganisation, conjugants split and a normal binary fission occurs restoring normal 1 MAC and 1 MIC (Siegel, 1963)



replication fidelity and a necessary adaptation when maintaining a separate germline (Sung et al., 2012).

There is no established methodology in *Paramecium* to transform the MIC genome so the only available reverse genetic methodology is that of gene knock-down with RNA interference (RNAi) (Marker et al., 2014). How-

ever, RNAi can be induced by one of two distinct but overlapping RNAi systems in *P. tetaurelia* (Marker et al., 2014). This is by microinjection⁴ of homologous transgenes (transgene-induced silencing) or by feeding *Paramecium* cells *Escherichia coli* transformed to produce sense and antisense transcripts for the target gene respectively (Galvani and Sperling, 2002) Furthermore, natural exogenous ssRNA in food bacteria of *P. tetaurelia* has been observed to produce low levels of silencing, therefore this mechanism is a likely a form of natural gene regulation used by *Paramecium* (Carradec et al., 2015). As *P. bursaria* shares the initial ancient WGD with *P. tetaurelia* (Aury et al., 2006) based on *P. tetaurelia* genes that are the product of this WGD it currently or previously will have possessed: a pair of Dicer/Dicer-like proteins, 6 pairs of Piwi genes and a single RdRP (Marker et al., 2014). Therefore, RNAi is likely available in *P. bursaria* as a means of testing predictions generated through transcriptomic and genomic investigation.

Paramecium appears to be particularly competent for endosymbioses with an array of over 60 genetically diverse putative endosymbionts described (Görtz and Fokin, 2009). This is no surprise as ciliates have been known to have bacterial (Görtz and Fokin, 2009), archael (Wrede et al., 2012) and eukaryotic (Kodama and Fujishima, 2009) endosymbionts. These endosymbionts range in degree from mutualist to parasitic and are cytoplasmic, endomicronucleic, endomacronucleic and/or perinuclear. As a serial phagotroph, *Paramecium* species are liable to infiltration by bacterial capable of escaping or resisting the phagosomal digestive process (?). The *Paramecium bursaria* micronuclei frequently contains bacterial endosymbionts. The closed nature of reproduction has been suggested as a reason why endonucleobioses are common in paramecium (Görtz and Fokin, 2009) Some endosymbionts exhibit high levels of adaptation, no longer able to be found free-living and with evidence of the genome reduction distinctive of endosymbiosis (Görtz and Fokin, 2009) The most frequently identified bacterial endosymbionts in German environmental samples are that of *Holospora caryophila*, *Holospora obtusa* and *Caedibacter caryophilus* Several of these endosymbionts have been shown to require specific *Paramecium* genes for maintenance Dohra et al. (1998).

1.3 PARAMECIUM BURSARIA - GREEN ALgal ENDOSYMBIOSES

The *Paramecium* - *Chlorella* endosymbiosis is established when *Chlorella* is phagocytosed by the serially phagotrophic *Paramecium* and is then able to escape the digestive vacuole. For this escape to take place, the endosymbiont must initially resist acidification caused by acidosome fusion with digestion vacuole. If the endosymbionts are able to resist this acidification they begin, through an unknown mechanism, to 'bud-off' from the initial phagosome into a new vacuole. This new perialgal vacuole (PV) is released into the cytoplasm and each PV contains an individual *Chlorella* cell (Kodama and Fujishima, 2009) The PV appears resistant to lysosome fusion and further digestive steps suggesting molecular modification of the vacuole membrane (Johnson, 2011) These perialgal vacuoles then bind the host cortex and compete for attachment with host structures known as trichocysts (Kodama and Fujishima, 2012) in a region with low to no lysosome activity (Kodama and Fujishima, 2009) This suggests

⁴This common RNAi pathway can be invoked by both Direct injection of dsRNA into the cytoplasm only triggers transient silencing, likely due to growth related dilution (Galvani and Sperling, 2002). Heritable silencing cannot be triggered by dsRNA (possibly due to insufficient RDR activity and absence of H3K9 histone factors in MIC) in *Paramecium tetaurelia* (Chalker et al., 2013)).

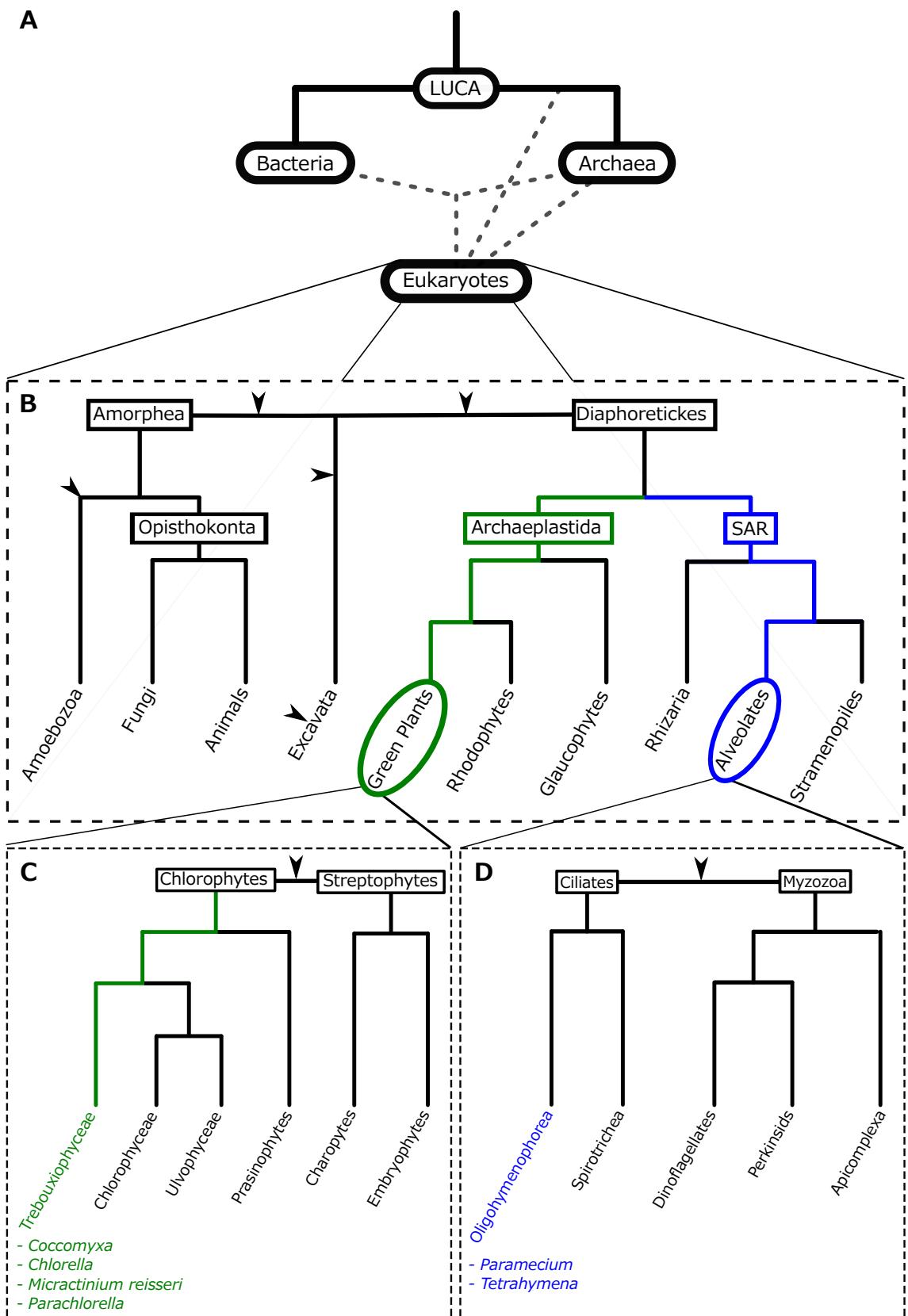


Figure 1.2.4: **A:** Schematic of the current best estimate of the tree of life demonstrating the 2D and 3D hypotheses, dashed lines indicate multiple potential branch location, arrowed lines demonstrate known endosymbiotic events (based on work reviewed in (Gribaldo et al., 2010)) **B:** Schematic of the current known eukaryotic portion of the tree of life (based on work reviewed in (Burki, 2014; Adl et al., 2013), **C:** Schematic of phylogeny of the ciliates (based on work by (Bachvaroff et al., 2011) showing Oligohymenophorea containing *Paramecium* and *Tetrahymena* and sister group Spirotrichea containing *Euplotes* and *Oxytricha*. **D:** Schematic of phylogeny of the green algae (based on work reviewed in (Leliaert et al., 2012)

the observed resistance to lysosome fusion may be a by-product of localisation. As few as a single algal cell can infect the host (Weis and Ayala, 1976) however, the majority of *Chlorella* are digested especially non-competent strains (Kodama et al., 2007). Furthermore, it has been established that *Chlorella* strains are fairly host-specific. For example, Summerer et al in 2007 (Summerer et al., 2007) showed that *Chlorella* isolated from other ciliates were able to establish endosymbioses with *P. bursaria* however, those isolated from cnidarian *Hydra* were not. This paper also showed *P. bursaria* favours its symbiotic partner over those isolated from other ciliates when given the choice, this suggests specific adaptations have taken place between host and endosymbiont (Summerer et al., 2007) Free-living *Chlorella* strains do rarely establish endosymbioses with *Paramecium* (Siegel and Karakashian, 1959), however they are generally only able to infect fewer *Paramecium* and establish much smaller endosymbiotic populations within the host than the symbiont strains (Siegel and Karakashian, 1959)

Once established, the symbiosis appears to be mutually beneficial with an observed flux of amino acids and CO₂ to the endosymbiont and oxygen and photosynthate (principally maltose) to the host as a function of light levels (Karakashian, 1963). The extent of this endosymbiosis is such that *Chlorella* is capable of supporting *Paramecium* in media without its typical bacterial food-stocks and conversely the *Paramecium* is capable of supporting the phototrophic *Chlorella* in the dark for ~2 weeks (or up to 51 endosymbiont cell divisions) suggesting considerable bi-directional nutrient flux (Siegel and Larison, 1960; Karakashian, 1963). It should be noted that for longer periods in the dark or when a bacteria-free culture is used in the dark the host will digest the endosymbionts (Parker, 1926) From an ecological perspective, this endosymbiosis can be considered as a means of acquired phototrophy (or mixotrophy), a tactic believed to be advantageous for survival in patchy oligotrophic environments by providing fixed carbon to cover respiration requirements (Putt, 1990). This is largely supported by studies, such as Karkashian's 1963 paper, showing that with a sufficient concentration of bacterial feedstock in the media the growth rate of asymbiotic *Paramecium* ('bleached') and *Paramecium* with *Chlorella* endosymbionts are largely equal. This threshold is estimated to lie between 10⁶ and 10⁷ bacteria per ml. However, as this is generally a much greater concentration than found in the natural environments of *P. bursaria* the endosymbiosis offers a considerable adaptive advantage to the host (Karakashian, 1963). As temporary acquisition of phototrophy is estimated by some research (Raven, 1997) to be less energetically costly than the permanent maintenance of plastids (via endosymbiosis or kleptoplasty) within the host this indicates that this endosymbiosis likely provides other host benefits beyond just the energetics of acquired phototrophy. These include:

- Exploitation of low oxygen environments by the host (as the photosynthesising endosymbiont is capable of providing oxygen to the host (Reisser, 1980)).
- Photoprotection and protection against 257nm and 282nm UV radiation potentially via endosymbiont pigmentation and localisation to shield host nuclei (Sommaruga and Sonntag, 2009; Summerer et al., 2009; Miwa, 2009). This is especially important as the AT-rich *Paramecium* genome is likely prone to UV-damage via the formation of cyclobutane thymine dimers (Sommaruga and Sonntag, 2009).
- Protection against predation (Berger, 1980). The exact mechanism by which this occurs is unknown, how-

ever, it has been observed that mixotrophic ciliates are able to move in rapid ‘jumping’ movements. This is hypothesised as being an energetically costly escape reaction made possible by sugar-rich photosynthate mixotrophic ciliates gain from their algal endosymbionts (Pérez et al., 1997). Intriguingly, this protection against predation occurs despite endosymbiont displacement of trichocysts (defensive cellular structures) for attachment to the ciliate cortex (Kodama and Fujishima, 2011).

- Protection against undesired endosymbionts and/or parasites. Algae in *P. bursaria* form an antagonistic relationship with some bacterial endosymbionts but there is experimental evidence that *P. bursaria* can only be infected by bacteria and yeasts after *Chlorella* is eliminated (Gortz, 1982). This is consistent with bacterial symbionts having been repeatedly identified as providing resistance to parasites in organisms such as the insects (Martinez et al., 2014).
- Protection against chemical toxins, for example symbiotic *Paramecium* have a much higher survival rate (96%) to 0.5 mM nickel chloride (NiCl_2) than asymbiotic *Paramecium* via an undetermined mechanism (Miwa, 2009)
- Increased thermotolerance (tested at 42°C) (Miwa, 2009), again, by unknown mechanisms but potentially related to the undefined means of perialgal vacuole attachment to the cell cortex.
- Protection against excessive oxidative burden (potentially due to endosymbiont dismutases and catalases) (Hörtnagl and Sommaruga, 2007) and hydrogen peroxide (hypothesised by Miwa as being due to the improved energetics of the symbiotic host) (Miwa, 2009).

In return, the endosymbiont also appears to gain several advantages including a generally much increased level of photosynthetic activity (Sommaruga and Sonntag, 2009):

- CO_2 from the host (Parker, 1926)
- Nitrogen supply (Johnson, 2011).
- Amino acids including L-glutamine (likely an important nitrogen source) and L-arginine, L-asparagine, L-serine, L-alanine and glycine (Kato and Imamura, 2009b).
- Host supplied divalent cations such as K^+ , Mg^{2+} , and Ca^{2+} . All of which have key roles in photosynthesis (Kato and Imamura, 2009b).
- Protection against *Paramecium bursaria* – *Chlorella* Virus (PBCV) (Yashchenko et al., 2012) a large isocahedral dsDNA, 330kbp virus with 133-genes that lyses symbiotic Chlorella when isolated from the host (Van Etten et al., 1983). This potentially occurs by preventing contact between PBCV and the endosymbiont.
- Effective photo-accumulation and increased mobility (Niess et al., 1982a).

This exchange of materials between host and endosymbiont is regulated by an effective biochemical ‘bartering’ system with numerous feedback cycles. For example, the release of endosymbiont photosynthate is dependent on

Ca^{2+} . This ion is provided by the host and also has a role in the up-regulation of photosynthesis (as proxied by oxygen evolution) (Kato and Imamura, 2009b). Once photosynthate is released into the PV lumen endosymbiont H^+ -ATPases are activated which allow the generation of the H^+ gradient necessary for endosymbiont uptake of host-provided amino acids via a set of amino acid-proton symporters (in the same manner as (Camoni et al., 2006)) (Kato and Imamura, 2009b). This proton gradient will potentially lead to further photosynthate release due to observed pH-dependence of this (Kato and Imamura, 2009b). As we can see the more photosynthate supplied to the PV lumen the greater the uptake of provided nitrogen sources. Intriguingly, from experiments using cycloheximide to selectively interrupt endosymbiont but not host protein synthesis it appears that the maltose exporter that is responsible for export of photosynthate from the PV lumen into the host cytoplasm is endosymbiont derived (Muscatine, 1967). However, unless photosynthesis is also inhibited (using DCMU) the build up of photosynthate without exportation in the PV triggers the swelling of the vacuole up to 25x its original size. This removes the vacuole from the region in which it is protected from lysosome fusion and leads to the digestion of the endosymbiont (Kodama and Fujishima, 2009). So, here we can see further regulation of the relationship – in which the endosymbiont is degraded if it does not release photosynthate to the host. This also demonstrates the importance of cytoplasmic localisation and the conditions in the PV to the control of this relationship.

On top of this system of secretion, uptake and feedback there have also been several other observed regulatory interactions between host and endosymbiont. The most apparent of these are the synchronising of cell division and circadian rhythms between host and endosymbiont. This integration is evidenced by the introduction of endosymbiotic *Chlorella* being sufficient to recover a circadian rhythm in arrhythmic *Paramecium* mutants (Miwa, 2009) This regulation of the timing of cell division for both members of the system appears well co-ordinated and takes place in such a way that neither host or endosymbionts outgrow one another (Kadono et al., 2004; Takahashi et al., 2007). The importance of regulation of endosymbiont distribution at host division is evidence in the only natural aposymbiotic *P. bursaria* mutant which has an impairment in this mechanism and thus can't maintain endosymbionts (Tonooka and Watanabe, 2002).

1.4 CONCLUSION

In conclusion, understanding the mechanisms by which primary and secondary photosynthetic endosymbioses have occurred is one of the most significant outstanding problems in understanding the evolution of the eukaryotes. *Paramecium bursaria* and its endosymbiosis with *Chlorella* offers a useful system to investigate secondary photosynthetic endosymbioses before metabolic co-dependence has become fixed. As both organisms seem highly prone to forming endosymbiotic relationships with multiple other organisms as a serial host and serial endosymbiont respectively it may be possible by identifying the key molecular components of their relationship to understand what factors contribute to such prolific utilisation of endosymbioses. Furthermore, while there is considerable supporting literature and many established methodological techniques for working on these organisms individually and in endosymbiosis there have been relatively scant efforts using the latest -omics techniques and reverse genetics such as RNAi. Considering the historical role both organisms have played independently in

our understanding of endosymbiosis⁵ it is perhaps apt that further insight may be gleaned by applying the latest modern techniques to interrogate their relationship.

Therefore, the main aims of this research are to assess the utility of *P. bursaria* and its algal endosymbionts as model for the study of the evolution of endosymbiosis. Specifically, are they metabolically, transcriptomically and/or genetically tractable? Is this system amenable to “single cell” analysis? Can *M. reisseri* be separated from its *P. bursaria* as has been observed in other *P. bursaria-Chlorella variabilis* strains? Finally, can RNAi be used to test hypotheses generated using these “omic” analyses?

⁵Margulis was strongly influenced and inspired by research conducted in organisms closely related to both *Paramecium bursaria* and *Micractinium reisseri*. Specifically, the discovery of Tracey Sonneborn of non-mendelian cytoplasmic inheritance in *Paramecium bursaria* (Sonneborn, 1950) and the multiple lines of evidence of the presence of DNA within the chloroplasts gleaned from several species of green algae related to *Micractinium reisseri* (*Spirogyra* (Stocking and Gifford Jr., 1959), *Chlamydomonas moewussii*, and *Chlorella ellipoidea* Ris and Plaut (1962)).

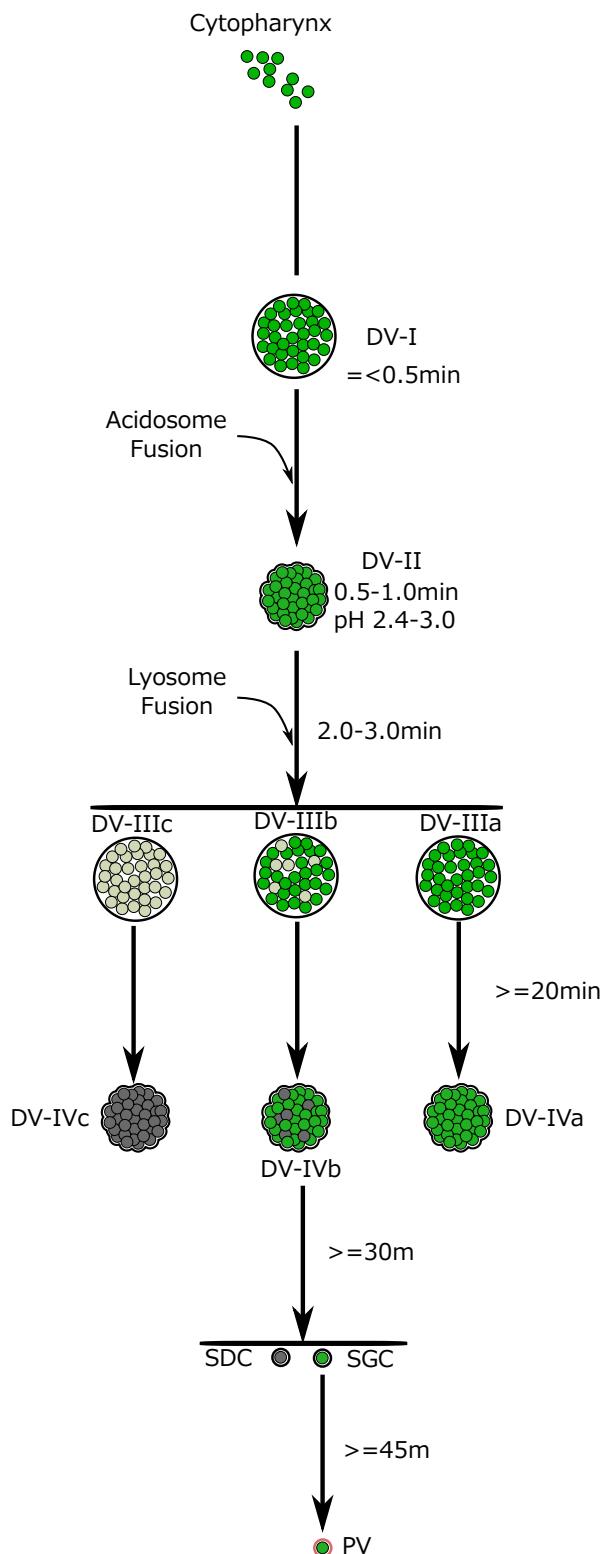


Figure 1.3.1: Process by which some endosymbiont escape digestion and generate perialgal vacuoles (PV). The digestive vacuole (DV) transitions through 8 distinct stages during PV differentiation. Initially a spherical DV-I is formed from the cytopharynx. Within a minute acidosomes fuse and differentiate this to the acidic DV-II. Primary lysosome fusion triggers swelling and conversion to DV-III. At this stage endosymbiosis capable endosymbionts escape digestion by an undefined mechanism. DV-III return to near neutral pH and are categorised into 3 categories based on them containing totally bleached, partially bleached or undigested algae. The DV-III condense into DV-IV following the same total, partial and undigested categories. Finally, DV-IVb undergo budding into a mixture of single digested *Chlorella* (SDC) and single green *Chlorella* (SGC). Finally, SGC translocate to the host cortex and further differentiate into PV (Kodama and Fujishima, 2009). Figure redrawn and modified from (Kodama and Fujishima, 2009).

Science is what we understand well enough to explain to a computer.

Art is everything else we do.

- Donald Knuth: *foreword to A = B by Petzovsek, Wilf and Zeilberger*

2

Methods

2.1 MICROBIOLOGY

2.1.1 STRAIN INFORMATION

During this project 3 *Paramecium bursaria* cultures have been used. These have been obtained from the UK Culture Collection of Algae and Protozoa (CCAP) and the Japanese National BioResource Project (NBRP). Specifically:

- CCAP 1660/12: *Paramecium bursaria* SW1 with *Micractinium reisseri* SW1-ZK ([Hoshina et al., 2010](#))
- CCAP 1660/13: *Paramecium bursaria* (unknown strain) with *Coccomyxa* CCAP 216/24¹
- NBRP Yad1g1N: *Paramecium bursaria* Yad1g with *Chlorella variabilis* 1N²

Both CCAP cultures (1660/12 and 1660/13) were isolated from the same pond in Cambridge, UK (pers. comm. Undine Achilles-Day CCAP, Oban, Scotland) CCAP 1660/12 was the principal culture and all genomic, transcriptomic and metabolomic analyses were conducted using these cultures. Theoretically, these 3 cultures provide us with *Paramecium bursaria* strains harbouring members of 3 of the 4 species of green algal *Paramecium* endosymbiont (see Chapter 3 for more details).

¹This is a mixed culture containing both CCAP 1660/12 strain with *Micractinium* and the *Coccomyxa* bearing strain, the *Coccomyxa* endosymbiont has been further isolated in CCAP under the description CCAP 216/24 (pers. comm. Undine Achilles-Day CCAP)

²Yad1g1N host is mating type 1 and was created by mixing of isolated and cultured endosymbiont (*Chlorella variabilis* Clone 1 (known as 1N strain))

2.1.2 MEDIA AND CULTURE CONDITIONS

All *P. bursaria* and green algae cultures were maintained in New Cereal Leaf-Prescott Liquid (NCL) media ($4.3\text{g/l } \text{CaCl}_2 \cdot 2\text{H}_2\text{O}$, 1.6g/l KCl , $5.1\text{g/l K}_2\text{HPO}_4$, $2.8\text{g/l MgSO}_4 \cdot 7\text{H}_2\text{O}$, 1g/l wheat bran , gravity filtered via GF/C paper and autoclaved) (CCAP, 2012) and stored in an incubator at 15°C with a 12:12 light:dark cycle. The incubator was lit using 2 21W 865 daylight fluorescent tubes, producing 2000 lumen each. Cultures were sub-cultured approximately every 2 weeks using fresh NCL media and were inspected using light microscopy to monitor health. No bacteria was added to cultures used prior to “omic” analyses but otherwise the medium was bacterised with *Klebsiella pneumoniae* SMC (strain donated by the Meyer Lab, Ecole Normale Supérieure, Paris, France) the day before use.

2.2 OMICS

“-omic” technologies are those aimed at globally characterising a class of biomolecules within a specific biological sample (characterising the “-ome”). The major areas of this are genomics, transcriptomics, metabolomics and proteomics. Genomics aims to characterise DNA and generally involves sequencing the genome, it is used to discover and describe genes (and non-coding DNA) and, by comparison with other genomic datasets, their evolution. Similarly, transcriptomics is orientated around the characterisation of the RNA present in a sample. This can include the canonical messenger RNA (mRNA) transcripts but also other RNA elements i.e. non-coding RNA (ncRNA) such as small interfering RNAs (siRNA) and micro RNAs (miRNA) and generally involves sequencing the RNA fraction of interest. Transcriptomics can be used to catalogue transcripts (and their variant splices), aid genome annotation, and/or assess transcriptional response to a given condition or cellular state (Wang et al., 2009). Metabolomics seeks, instead, to identify and quantify small biomolecules that make up the terminal and intermediate products of cellular metabolism e.g. carbohydrates, alcohols, and amino acids. Finally, proteomics characterises the proteins present in a sample. Typically, the metabolome and proteome are interrogated using various forms of mass-spectrometry. There are also a plethora of additional approaches which seek to characterise different subsets of these biomolecules e.g. epigenomics (epigenetic modification to DNA such as methylation and histone binding), glycomics (characterisation of cellular saccharides). “Meta-...-omics” is the application of specific “omic” method to a biological sample containing multiple organisms. For example, “metagenomics” has been used to investigate the cellular community composition of marine micro-eukaryotes (Cuvelier et al., 2010) and “metatranscriptomics” has been used to analyse the transcriptomes of the microbes present in the gut of metazoa (Perez-Cobas et al., 2013).

The utility of “-omic” approaches is they allow a researcher to characterise a high proportion of a biological system’s function in a way that is faster, cheaper and requires less *a priori* knowledge of the system than more targeted approaches. For example, in order to estimate the abundance of all mRNA transcripts in a sample using specific approach such as RT-PCR would require sequence knowledge to design primers as well as an infeasible amount of reactions to acquire a characterisation comparable to that obtainable by a transcriptomic approach such as RNA-Seq. Additionally, due to being “non-targeted” (or rather less targeted) “omics” also removes one aspect of researched-induced bias caused by a conscious selection of molecule specific probes. By not considering elements

of a system in isolation like the classic methodologically reductionist³ approaches “omics” can reveal complex systemic mechanisms/features (or at the extreme “emergent properties”) that would otherwise have been missed (Fang and Casadevall, 2011). Experimental design is very important when using of “-omic” platforms as the the number of biological replicates tends to be far smaller than the number of parameters/metabolites/transcripts being studied.

However, until relatively recently “omic” methodologies were restricted to specialised institutions and well characterised “model” organisms. While, *Paramecium bursaria* and green algae such as *Micractinium reisseri* could be considered “model” organisms throughout the early days of molecular biology, they are much less frequently studied in the genomics era (2000-today) particularly compared to organisms such as *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. Fortunately, due to the development and maturation of both technologies and databases the potential for functional and adaptive analysis of non-model organisms using combined “omics” (i.e. using genomics as a reference to guide subsequent transcriptomics) approaches (e.g. (Muñoz Mérida et al., 2013; Feldmesser et al., 2014)) has recently been demonstrated. Additionally, there are two other developments which make *P. bursaria*-*M. reisseri* (*PbMr*) amenable to “omic” analysis: *de novo* transcriptomics, which dispense with the need to generate accurate genomes in the relatively genetically intractable *Paramecium* (e.g. (Kodama et al., 2014)), and single-cell approaches which allow fine-grained analysis of the *Paramecium bursaria* - green algal relationship on a cell-by-cell basis.

It should be noted that care must be taken with “omics” approaches as they can easily become purely descriptive, and at worse generate models that lack any biological relevance (Fang and Casadevall, 2011). This concern holds for all systems-level approaches and has been frequently raised and discussed in the context of genomics (Dougherty, 2008). Therefore, it is crucial to supplement “omic” approaches with targeted methods in a way that compensates for the weakness of each type of method. Specifically, the systems approach should be used to generate novel and interesting hypotheses which can then be tested in isolation using reductionist methods (Casadevall and Fang, 2008). For example, “omics” methods could be used to create a model of inter-organism host-endosymbiont metabolism and targeted approaches such as RNAi could then be used to test hypotheses generated by this model i.e. testing that a particular transporter protein is responsible for the transfer of metabolites by knocking out that transporter and observing the resultant phenotype: is the relationship perturbed in a predictable manner?

2.2.1 GENOMICS AND TRANSCRIPTOMICS

2.2.1.1 DNA SEQUENCING

In the majority of cases, genomics and transcriptomics are both synonymous with the sequencing nucleic acids. Earlier approaches, based upon the fluorescent marking of the hybridisation of DNA and/or RNA to arrays of short complementary probes e.g. genomic tiling arrays and the transcriptome microarrays (Mockler and Ecker,

³Epistemological reductionism: “explain all biology in terms of physics and chemistry” (Crick, 1966) i.e. biology is applied chemistry which is applied physics which is applied maths. Ontological reductionism: a biological system is only the sum total of its component molecules and their interactions. Methodological reductionism: examination of simple components can be used to understand complex system (Fang and Casadevall, 2011)

2005), are of more limited utility. Relative to sequencing-based approaches these methods require relatively more prior knowledge of the organism and require a custom array to be designed for any novel system. Additionally, while microarrays can determine relative expression levels of transcripts by the comparison of the fluorescence intensity at given complementary probe(s) the continuous nature of this output, difficulty distinguishing alternative isoforms and more limited dynamic range (combined with previously mentioned limitations) has meant the sequencing of cellular transcripts (RNA-Seq) has largely supplanted microarrays (Wang et al., 2009). However, both SNP tiling arrays and microarrays do have the advantage of throughput and ease of analysis in situations where the host organism is well known and suitable arrays have already been designed and evaluated. For this reason they are still frequently encountered in specialist area of medical diagnostics.

While it is possible to directly sequence RNA transcripts (Ozsolak et al., 2009) most approaches first utilise a reverse transcription (RT) step to convert transcripts to cDNA. As ribosomal RNA makes up a sizeable proportion of RNA in the cell it is often necessary to enrich or select the RNA fraction of choice in order to minimise wasted effort when sequencing (Wilhelm and Landry, 2009). For eukaryotic mRNA enrichment this can be easily achieved by using poly-T primers during RT which selectively bind to the polyadenylated tail of these messenger transcripts. However, for bacterial/archaeal work and transcriptomic analyses focusing on non-polyadenylated transcripts such as ncRNAs/siRNAs/miRNAs etc. ribosomal depletion is used (O'Neil et al., 2013). This is a process by which ribosomal probes are attached to magnetic beads. Ribosomal RNAs bind to these probes and the magnetic beads can be used to partition the majority of ribosomal sequences away from the other RNA (O'Neil et al., 2013). This means that transcripts can be sequenced using the same methods and platforms as any other DNA sample with analysis only diverging against post-sequencing. It should be noted that there are potential disadvantages to this reverse transcription step and it can potentially generate artefacts and biases in the analysis (as well as placing limitations on the quality and quantity of input RNA) (Ozsolak and Milos, 2011) however, the advantages of the more developed DNA sequencing technology outweighs these disadvantages.

These DNA sequencing technologies can largely be divided into 3 technological eras with today (2015) broadly at the transition between 2nd and 3rd generations.

1st generation (also known as Sanger) sequencing technology originated in 1970s with the work of Sanger & Coulson (Sanger and Coulson, 1975; Sanger et al., 1977a,b) which developed sequence determination via the principle of chain termination during synthesis and subsequent determination of relative fragment sizes. Briefly, by having 4 separate reactions in which DNA synthesis terminates on the incorporation of dideoxy nucleotides (ddNTP) corresponding to each of the 4 principal DNA bases (i.e. ddATP, ddGTP etc.) you can generate a series of DNA fragments of various sizes. Size fraction separation of these fragments via methods such as gel electrophoresis means the DNA sequence can be easily read from the fragment size distribution across the 4 ddNTP reactions (Sanger et al., 1977b). This technique was used to sequence the first DNA genome (bacteriophage ϕ X174 (Sanger et al., 1977a)). The methodology was subsequently improved by use of fluorescently labelled ddNTPs by Leroy Hood, massively simplifying automation of the process (Smith et al., 1985, 1986). Further improvements followed throughout the 1990s and early 2000s such as capillary electrophoresis and other general throughput and length enhancements (Bonetta, 2006). Transcriptomic analysis was possible using Sanger sequencing by generat-

Library preparation

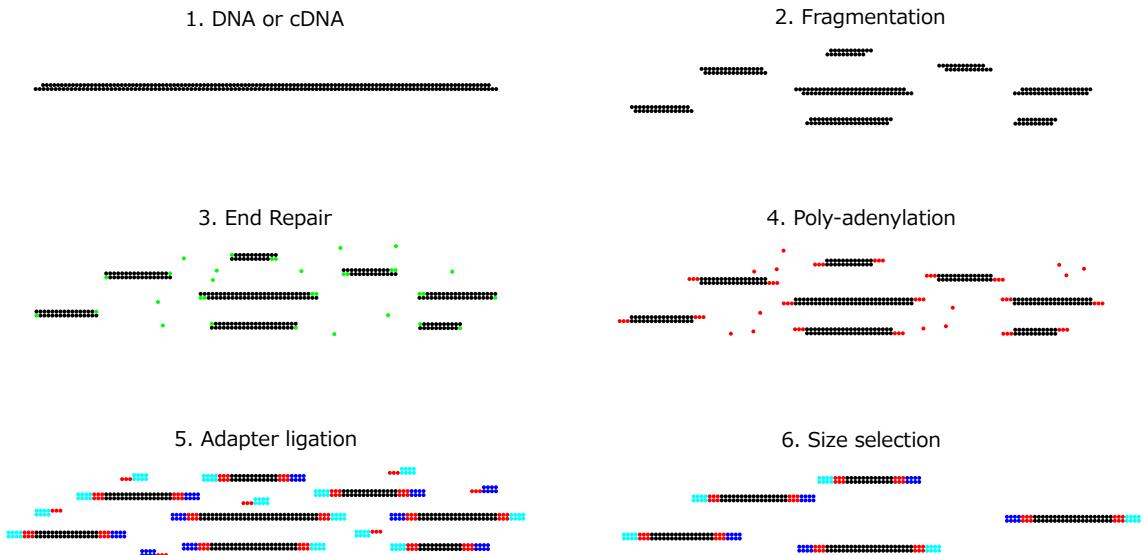


Figure 2.2.1: A brief overview of library preparation for Illumina modified from ([Mardis, 2008](#)) and Illumina TruSeq kit documentation

ing clone libraries from partial or complete cDNA and randomly sequencing clones ([Adams et al., 1991](#); [Gerhard et al., 2004](#)). However, while this did allow resolution of different isoforms and could be used to aid annotation ([Adams et al., 1991](#)) it was not possible to investigate relative expression levels beyond a broad identification of highly expressed transcripts based on the proportion of the cDNA/EST library they made up. Sanger sequencing's main utility lies in high quality short fragment (300 – 1000bp) sequencing to determine or confirm the sequence of specific DNA fragments such as vectors or PCR products ([Bonetta, 2006](#); [Tsiatis et al., 2010](#)).

2nd generation sequencing emerged commercially in 2005 with the work of both George Church and 454 Life Sciences ([Margulies et al., 2005](#)) and featured reduced individual reaction volumes, greater parallelisation (and so higher throughput), cell-free preparation without the need for time-consuming cloning of DNA fragments into bacterial vectors to generate clonal templates for sequencing, and direct sequencing detection obviating the need for size fractionation ([Jaszczyzyn et al., 2014](#)). These technologies generate huge amounts (on the order of 10^6 – 10^9 of relatively short (on the order of 10¹ – 10³bp) DNA sequences (reads) randomly sampled from the input (c)DNA.

Commercially available 2nd generation platforms include 454's GSFLX and GSJunior (now Roche), Ion Torrent's (now Life Technologies) PGM, Applied Biosystem's (now Life Technologies) SOLiD and Illumina's (formerly Solexa) HiSeq, MiSeq and older Gene Analyzer II ([Nederbragt, 2013](#)).

Although these platforms use a range of different implementations and tend to exhibit various different trade-offs (mainly in terms of number of reads and their respective lengths) they all largely follow the same basic process ([Shendure and Ji, 2008](#)):

1. Library generation: Randomly fragmenting input DNA into short fragments of a specific size followed by ligation of adapter sequences with some platforms allowing development of “paired-end” or “mate-pair” libraries in which each end of a fragment is sequenced separated with a known size unsequenced fragment

Paired-End Illumina Sequencing

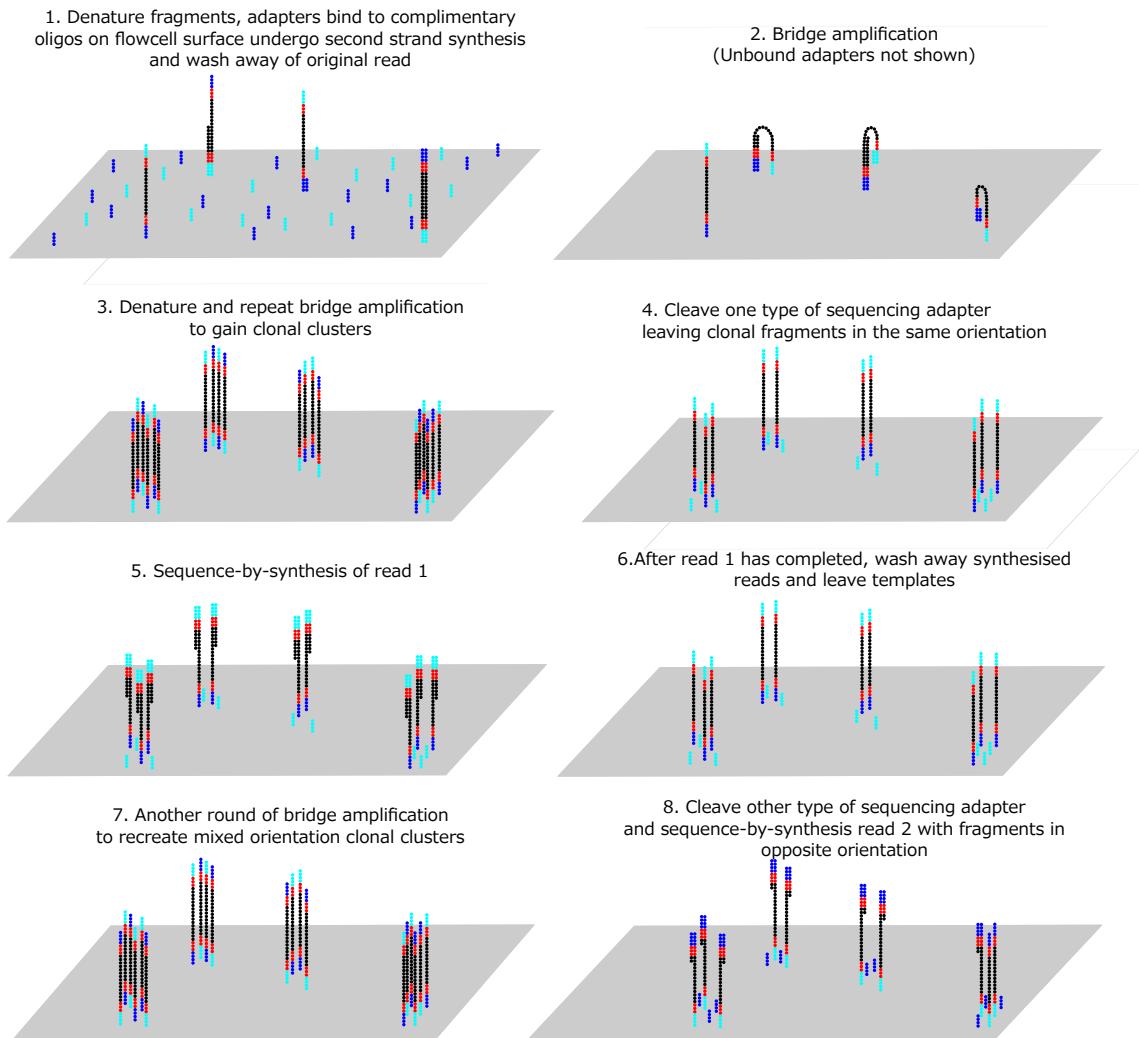


Figure 2.2.2: A brief overview of paired end sequencing in an Illumina flowcell after library preparation, derived from (Mardis, 2008) and Illumina

aiding subsequent assembly (see 2.2.1)

2. Clonal amplification: Generation of clonally identical spatially distinct clusters of DNA mainly via emulsion PCR (Dressman et al., 2003) (SOLiD, Ion Torrent, 454) or bridge PCR (Adessi et al., 2000; Fedurco et al., 2006) (Illumina) (see 2.2.2)
3. Sequencing-by-synthesis: In which a complementary DNA strand is generated base by base via sequentially flooding and clearing a chamber with each dNTP and a polymerase (or ligase in the case of SOLiD). On incorporation of a base into a cluster a detectable signal is released such as emission of certain wavelengths of light detectable using optics (e.g. Illumina, 454, SOLiD) or release of hydrogen ion (e.g. Ion Torrent).

The explosion in sequencing throughput on 2nd-generation platforms has driven a massive decrease in per-base sequencing cost and the subsequent expansions in the amount of available data (e.g. the US National Center for Biotechnology (NCBI)'s short-read archive (SRA)) has made both genomic and RNA-Seq analysis and annotation easier and more effective.

While, 2nd generation sequencing has driven down per-base sequencing costs the cost of library preparation has fallen more slowly (Blainey, 2013). For this reason, combined with the higher throughput it has become common to multiplex different samples during sequencing runs. Multiple distinct samples can be sequenced in the same reaction (e.g. flowcell lane for Illumina platforms) by adding an indexed tags during library preparation. These tags can then be used to partition the reads back to their original separate samples after sequencing.

The current *de facto* standard in 2nd generation sequencing is that of the bridge amplification based (Shendure and Ji, 2008) Illumina platforms (Regalado, 2014) due to relatively low error rate ($\leq 0.1\%$ (Glenn, 2011)), very high throughput (e.g. HiSeq2500 generates up to 400M 125 bp reads per run (1TBase of data) (Nederbragt, 2013)) and the lowest cost per Mb ($\leq \$0.04$ (Glenn, 2011)).

Finally, 3rd generation technologies are generally known as single-molecule sequencing. These platforms sequence individual DNA (or RNA molecules (Ozsolak et al., 2009)) without bias and error-prone amplification. The first 3rd generation platform was that of the now defunct Helicos Bioscience's Helicoscope (Harris et al., 2008) based on breakthroughs in the resolution of fluorescence visualisation using paired FRET methods (Braslavsky et al., 2003). There is only one publicly available platform: Pacific Biosciences (PacBio) RS platform. PacBio operates on a similar principle of sequencing-by-synthesis as the 2nd generation platforms but uses fixed polymerases at the base of specially wave-guide structures allowing the detection of fluorescence from a single reaction instead of many parallel reactions in a clonal cluster. This produces few (compared to 2nd generation platforms) long (20kb and longer) reads but has a high cost and high error rate (14%) (Jaszczyszyn et al., 2014) Another platform, currently in testing, Oxford Nanopore's MinION, reads individual strands of DNA through an array of pore proteins and determines the sequence at each pore based on the physical properties (impedance) of a particular set of bases.

Unfortunately, partly as an element of their relatively nascence and partly due to the poorer signal:noise of single molecule approaches compared to analysing large batches of identical DNA sequences, 3rd generation technologies have a relatively high error rate (Quail et al., 2012). Therefore, they are generally inadequate for eukaryotic assembly tasks in and of themselves (although that is changing as the platforms mature (?)). Where they have shown great utility is in conjunction with 2nd generation datasets as a scaffolding tool i.e. producing long noisy reads upon which more accurate but shorter reads can be assembled (Zhou et al., 2009). 3rd generation platforms are also highly useful in the resolution of structural variation and genomic repeats (Pendleton et al., 2015).

Therefore, all genomic and transcriptomic sequencing in this PhD has been performed using the 2nd generation Illumina GAII and HiSeq platforms due to their relative maturity, high-throughput, relatively accurate paired-end output making it currently the most amenable platform to effectively use *de novo* genomic and transcriptomics approaches. Additionally, Sanger sequencing has been used when accurate targeted sequencing was called for, such as investigating the taxonomic distribution of *Paramecium* green algal endosymbionts (see Chapter 3).

2.2.1.2 READ PRE-PROCESSING

All sequencing platforms involve base-calling procedures in which the continuous primary sequencing data e.g. fluorescence with Illumina or current with IonTorrent is converted into a discrete nucleotide sequence. This is typically achieved using parametric modelling and machine learning approaches optimised by the platform manufacturer ([Ledergerber and Dessimoz, 2011](#)). Researchers using mature platforms such as Illumina typically rely on the default base-calling. However, there is much active research in improving these systems particularly for SMRT sequencing. These include various advanced signal processing and machine learning methods such as Conditional Random Fields and Recursive Neural Networks.

The major 2nd and 3rd generation base-calling procedures output a FastQ formatted file. These are files containing the sequence of a read and a per-base quality score known as a Q score. These scores (Q) are calculated as a logarithmic relation to the base-called error probability (P)

$$Q = -10 \log_{10} P$$

and conversely:

$$P = 10^{-\frac{Q}{10}}$$

Therefore

$$Q = 30$$

corresponds to 99.9% base call accuracy

Before assembly, there are several stages of read pre-processing that generally takes place for both genomic and transcriptomic studies. Typically, there are 4 key steps:

- Library quality control and contamination screening
- Trimming sequencing adapters and low probability reads
- Error corrections
- Digital normalisation

Library quality control (QC) and contamination screening is frequently done by as part of the standard diagnostics of a sequencing facility. It is a highly useful step which allows investigation of potential technical problems with a given sequencing run. For this reason, QC is typically done by the sequencing service or facility operating the sequencers. Problems with library preparation and sequencing can be detected by analysis of things like the distribution of per-base quality scores across sequences, and the presence of massively over-represented K-mers in the raw libraries. The most commonly used quality control tool is FastQC ([Andrews, 2015](#)) which performs tests and visualises various library metrics such as the distribution of quality scores across reads, distribution of low quality base reads across the flowcell and the identification of over-represented K-mers and sequences in the library.

Identification of potential library problems can be used to either attempt to correct these problems, for example, if there are consistently low quality bases at the end of the reads then a harsher 3' trimming step can be used. Alternatively, libraries with many problems may require re-sequencing. Contamination screening is a highly important part of sequencing, especially for single cell approaches (see Chapter 4 for an in-depth analysis) as contaminant reads from non-target organisms can greatly complicate analysis and assembly of data.

While second-generation sequencing technologies have massively increased throughput on an individual read basis they exhibit a much higher error rate than earlier Sanger approaches with Illumina HiSeq reads showing 0.5–2.5% error rate (Kelley et al., 2010). Typically, Illumina errors are substitution errors (Yang et al., 2013a) and are distributed non-randomly across the read. The error rate increases from 5' to the 3' end of a given read (Liu et al., 2012). The presence of sequence error does lead to assembly error (Macmanes and Eisen, 2013). Sequencing error greatly complicates assembly graphs and increases the computational demands of graph traversal. Therefore, before assembly this error must be minimised. This generally involves two processes, read trimming and explicit error correction.

Read trimming serves two roles: the removal of contaminating sequencing adapters that may have infiltrated the library during sequencing, and the removal of low quality sequencing data. Low quality sequence data is removed as these reads/parts of reads are more likely to contain errors. There are many available tools for read-trimming e.g. Trimmomatic (Bolger et al., 2014), Sickle (Joshi and Fass, 2015), FASTX-toolkit (Gordon and Hannon, 2010), PRINSEQ (Schmieder and Edwards, 2011) and cutadapt (Martin, 2011). These generally fall into two algorithmic groups: running-sum based approaches e.g. Cutadapt and ERNE-FILTER and window-based e.g. FASTX, PRINSEQ, Sickle and Trimmomatic (Del Fabbro et al., 2013). Briefly, running-sum approaches involve the calculation of a score for each base, i , from 3' to 5' e.g. $s(i) = s(i + 1) + \text{quality}(i) - Q$ for a minimum quality threshold Q . These s are used to determine the start of the trimming point. Alternatively, window-based approaches generally calculate the average quality score across a sliding-window of N bases. If the score in a given window drops below the minimum average quality it is then trimmed.

Currently, there is no clear answer to the question: which is the best trimming algorithm. This is due to many dataset-specific effects as well parameter-dependence (Del Fabbro et al., 2013). In general, most trimming tools have been found to largely perform equivalently across multiple RNA-Seq and DNA-Seq datasets and applications (see File S2 (Del Fabbro et al., 2013)). Due to its ease of use and maintenance of paired-read correspondance, Trimmomatic has been the primarily used trimming tool throughout this thesis.

Traditionally, most projects use conservative trims which only accept reads and bases above a high threshold such as an average score of Q_{30} (e.g. (Looso et al., 2013)). However, recently, there have been empirical studies suggesting the optimal approach is permissive trimming (e.g. $\geq Q_2$) followed by explicit error correction (MacManes, 2014).

Error correction of Illumina sequencing reads has been acknowledged as an increasingly important step in the creation of both high quality genomic (Schatz et al., 2012) and transcriptomic (Macmanes and Eisen, 2013) assemblies. This pre-process step generally operates on the assumption that errors are infrequent and random. An error at a given position can thus be identified and corrected by comparison to the sequences of the other reads

which are sample from the same region (Yang et al., 2013b). Specifically, if the majority of other reads feature the same sequence but a single read has a single substitution this is more likely to be product of sequencing error than biological diversity.

This is most typically achieved using probabilistic approaches after decomposition of reads into K sized substrings (known as K-mers). These K-mers are then analysed via spectral techniques and the construction of hamming graphs (e.g. (Nikolenko et al., 2013)) to identify related low-abundance and likely erroneous K-mers. Some approaches will also integrate sequence alignment and quality score features. However, K-mer approaches are generally more computationally efficient and more effective at removing sequencing error than quality score based approaches (Zhang et al., 2014). See (Molnar and Ilie, 2014) for a review of error correction algorithms.

Error correction is more difficult for transcriptomic datasets (and single cell data) because the assumption of uniform coverage is not true for these datatypes (Macmanes, 2015). Therefore, correction of these datasets generally relies on explicitly probabilistic approaches that avoid this assumption by analysing relative abundances and methods such as bayesian subclustering of the hamming graphs (Nikolenko et al., 2013).

The final form of read-processing is arguably a variant of error correction, it is known as digital normalisation (Brown et al., 2012). This simply involves the elimination of redundant read data in a given library. As short-read sequencing involves the random sampling of the transcriptome or genome there is a high level of redundant sequencing where many reads are derived from the same template. The number of reads that map to a given portion of a template is known as the coverage. While the high levels of coverage are necessary for accurate assembly reconstruction it generates a computational burden and increases the problem of sequencing error. Therefore, by normalising coverage by progressively filtering out the most abundant reads it is possible to minimise sampling variation and generate much smaller libraries that still contain nearly the same amount of information (Brown et al., 2012). This means the computational demands of assembly are much lower and thus it is easier to tune assembly parameters.

2.2.1.3 ASSEMBLY

There are two main approaches to both genome and transcriptomic assembly - referenced and *de novo*. A referenced assembly consists of the alignment of processed reads to a prior assembly or reference using specialised short-read aligners such as Bowtie or BWA.

On the other hand, *De novo* assembly algorithms form two main groups: Overlap-Layout-Consensus (OLC) methods and de Bruijn graph methods. OLC are conceptually relatively simple, a graph is constructed based on the overlap of sequencing reads determined using standard pair-wise alignment algorithms. Each graph vertex represents a read and an edge connecting a pair of vertices indicates overlap between those reads.

generally falls into two major groups of graph algorithms.

While explained later in ??, generally OCC assembly is reliant on fast and accurate overlap calculations while de Bruijn approaches require robust error correction (Palmer et al., 2010).

mapping reads to reference can introduce biases even with relatively (Wang et al., 2012) but even if the reference is divergent it may still produce a higher quality assembly than *de novo* methods (Vijay et al., 2013) Soap-

denovo trans and trinity performed better than transabyss (Vijay et al., 2013) In De Bruijn graph assembly generated from exact subsequence matches - errors exponentially increase possible graph traversal paths increasing graph complexity and assembly time and memory (Pop, 2009)

<https://www.biostars.org/p/52178/> sequencing papers theory Spades (Bankevich et al., 2012)

Trinity (Grabherr et al., 2011)

2.2.1.4 THE PROBLEM WITH PLOIDY

One important complication in the assembly of eukaryotic genomes relative to bacteria or archael sources is the issue of highly heterozygous polyploid genomes. This is problematic as the de Bruijn graphs constructed during assembly rapidly increase in complexity when reads from heterozygous samples become incorporated (Kajitani et al., 2014). This is because k-mers derived from heterozygous regions of homologous chromosomes will partition the assembly graph into bubbles that cannot be easily or accurately resolved by most assemblers expecting only limited structural variation during assembly. Previously, attempts to work around this included inbreeding to generate homozygous lines or fosmid based approaches.

Specialised genome assemblers have been produced to address this problem, for example Platanus performed well on both highly and lowly heterozygous genomes using k-mer autoextension approaches and by merging haplotypes at both contig assembly and scaffolding steps, as well as incorporation of various heuristics involving bubble resolution (Bradnam et al., 2013; ?)

Likewise, transcriptome assembly complexity rapidly increases with the number of alleles expected per gene determined by ploidy, heterozygosity and complex gene families and in turn how many transcripts per allele in light of alternative splicing.

This is particularly problematic in the PbMr system owing to the massive ploidy of the host *Paramecium bursaria* and the numerous whole genome duplications in its relatively recent evolutionary history (1). This also explains the difficulties in using sister species, as the most sequenced Paramecium genus species are the aurelia complex which have undergone 2 WGD since divergence with *P. bursaria* (McGrath et al., 2014).

2.2.2 METABOLOMICS

Metabolomics is based around the detection and characterisation of cellular metabolites via a range of techniques. By qualitatively and/or quantitatively analysing these molecular substrates, cofactors and products of cellular metabolism a researcher can determine cellular state or response at the level. It is a highly useful set of techniques as metabolites are a direct signature of biological activity. This means metabolomics data is the final level in the path that links genotype to cellular phenotype (Fiehn, 2002).

Metabolomics can range from target analysis (focussing on a single metabolite) to metabolic profiling (orientated on the metabolites of a specific pathway or type) to full metabolomics (untargeted analysis of the entire metabolome) to metabolic fingerprinting (high throughput metabolomics of a large number of organisms, also sometimes referred metabolomics) (Fiehn, 2002). Metabolomics is conducted using a variety of analytical chem-

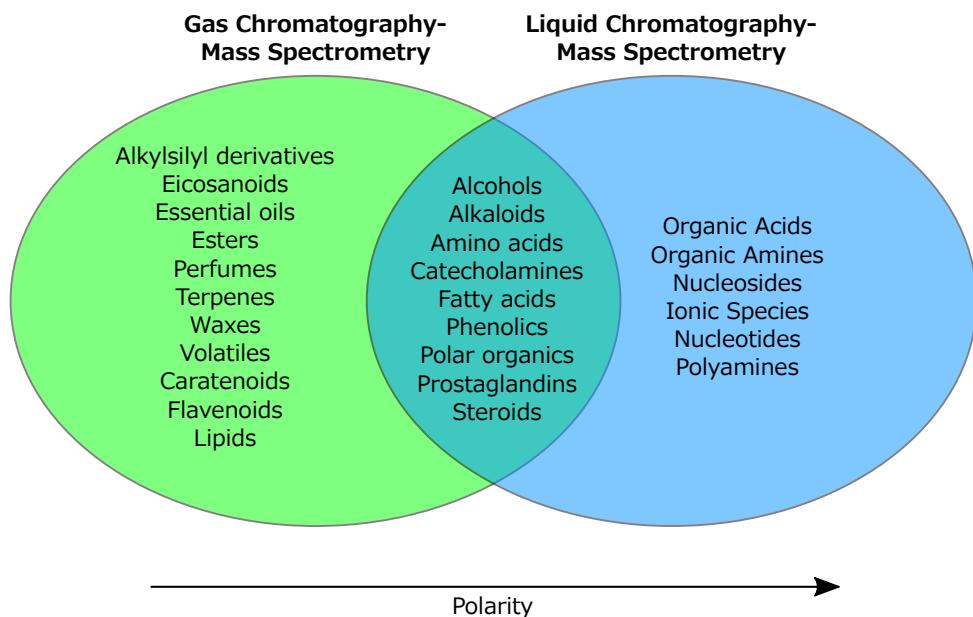


Figure 2.2.3: Groups of metabolites by which separation method is best suited for their analysis. Figure redrawn from <https://www.agilent.com/cs/library/selectionguide/Public/s5989-6328EN.pdf>

istry methods including nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), light (usually infrared (IR) or ultraviolet (UV)) spectroscopy (Kafsack and Llinás, 2010).

The difficulty of mass spectrometry is using a method that allows quantification and analysis of the diverse range of metabolites present in living cells. Due to the large number of varying physico-chemical properties and a broad dynamic range of concentrations no one single metabolomic method will fully characterise all metabolites with equal accuracy. Therefore, a comprehensive analysis of the metabolome is reliant on both careful experimental design and the use of a range of analytical techniques (Dettmer et al., 2007). Specifically, in the case of MS based methods, multiple separation methods must be used i.e. Gas (GC) and Liquid Chromatography (LC) (fig. 2.2.3) as well as a range of capture parameters.

Mass spectrometry works on the principle that a charged particle moving through an electromagnetic field is subject to the following two laws: the Lorentz force law (??) and Newton's second law of motion (??)

$$F = Q(E + \nu \times B)$$

$$F = ma = m \frac{dv}{dt}$$

where F is the force applied to the ion, m is the mass, a the acceleration, Q the electric charge, E the electric field and $\nu \times B$ the cross product of the ion's velocity and magnetic flux density.

Therefore, by combining these two identities:

$$\left(\frac{m}{Q}\right)a = E + \nu \times B$$

m/z denotes dimensionless quantityy from dividing the mass number of the ion by its charge number.

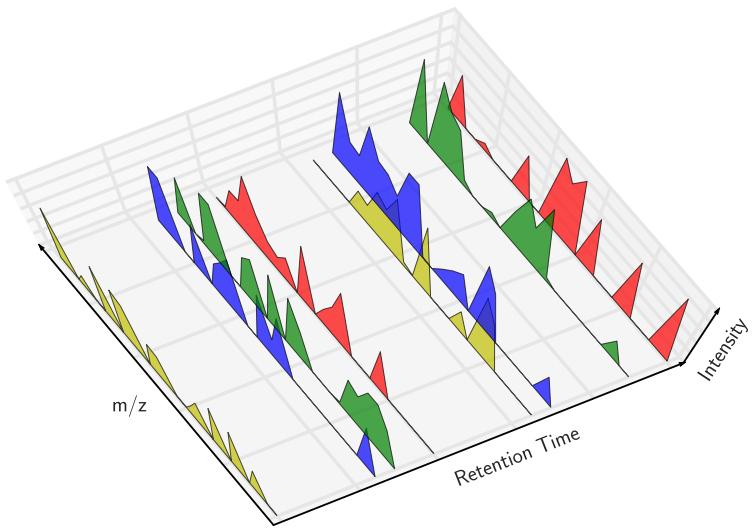


Figure 2.2.4: Visualisation of Mass Spectrometry data when coupled with chromatographic techniques. The X axis represents the m/z ratio of an ion, the Y axis the retention time within the chromatographic separation method (GC, HPLC etc.) and Z the intensity detected at a given m/z

Mass spectrometry involves 4 components with numerous options and alternative methods for each optimised for different analytes:

- Sample introduction
- Ion source
- Mass analyser
- Mass detector

As mass analysers are only capable of analysing charged ions in a gaseous phase it is necessary to volatilise and charge analytes during sample introduction and ionisation respectively.

While samples can be directly introduced and vaporised using heat or similar, typically most metabolomic analyses involve the chromatographic separation of samples using either liquid (LC) or gas (GC) chromatography.

2.3 MACHINE LEARNING AND STATISTICAL PATTERN RECOGNITION

Machine learning is a field of computer science devoted to the challenge of developing and applying algorithms capable of automatically inferring and utilising patterns in data (Murphy, 2012). A commonly used formal definition of machine learning: “A computer is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E.” (Mitchell,

1997). ML encompasses techniques and methods from various areas including statistics, optimisation/control engineering, neuroscience and artificial intelligence. Applications range in complexity from simple linear regression to deep convolutional neural networks with millions of free parameters running on dedicated super-computers (Wu et al., 2015) which are capable of beating human-performance on complex image classification tasks (e.g. IMAGENET (Russakovsky et al., 2014; He et al., 2015)).

Typically, we seek to set the parameters (" θ ") of a function in such a way that another property is minimised. For example, in linear regression the aim is to find parameters of a straight line $h_{\theta}(x) = \theta_0 + \theta_1 * x_1$ which minimise the distance between the line and the data (for example, the sum of squares distance). This distance/error is calculated using something known as the cost function e.g. $J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$ (where m is the number of x, y pairs in the dataset for linear regression). Most algorithms will seek to minimise the value of this cost function $J(\theta)$ with respect to the parameters of the original function $h_{\theta}(x)$. Typically, this is achieved using a variety of algorithmic optimisation techniques. The most prevalent of these are gradient descent based methods in which the value of θ is modified in the direction of the gradient of the cost function (determined using the partial derivative of J with respect to θ : $\frac{\partial J_{\theta}}{\partial \theta}$).

In an ideal world, the best machine learning model trained using our data will generalise well for novel data generated from the same underlying process which generated the training data. This is known as generalisability and it plays into the concept of 'fit'. A model that minimises its particular cost function on the training dataset has been fit to that dataset, however, it is possible for the model to fit the training data in such a way that it has low error on the training data but performs incredibly poorly when applied to new data from the same process. This is typically the case when a model has overfitted the data.

The classic example of this is fitting a line to a set of points using a high degree polynomial (see fig. 2.3.1). This polynomial will perfectly pass through all the points but is likely to be a worse predictor for the value of some new data than a much simpler model that while it fits the original training data, may not fit quite as well. Likewise, a model that is misspecified or cannot fit the training data well e.g. the training data follows a non-linear distribution but the model is linear, is known as underfitted. Underfitted models will perform poorly on both the training and test data. However, it isn't particularly useful to only discover how useful your model is likely to be on the test data therefore almost all machine learning analyses uses the principle of cross-validation. Cross-validation is the partitioning of the training dataset to create a validation dataset which can be used as a proxy test set.

Unfortunately, no single model will perform best for all tasks (to paraphrase and simplify Wolpert and McCreedy's "No Free Lunch Theorem" (Wolpert, 1996)), there are no shortcuts in machine learning (and many other areas) or optimisation. Therefore, testing different models (and hyperparameter values) using cross-validation is key to generating a useful model. Another important way to prevent overfitting is to introduce regularisation in the cost function, in other words a term which penalises model complexity.

Machine learning is typically divided into 2 main subsets depending on the nature of the dataset involved: supervised learning (e.g. classification and regression) and unsupervised learning (e.g. clustering, density estimation and dimensionality reduction). There are also approaches that blend features of both supervised and

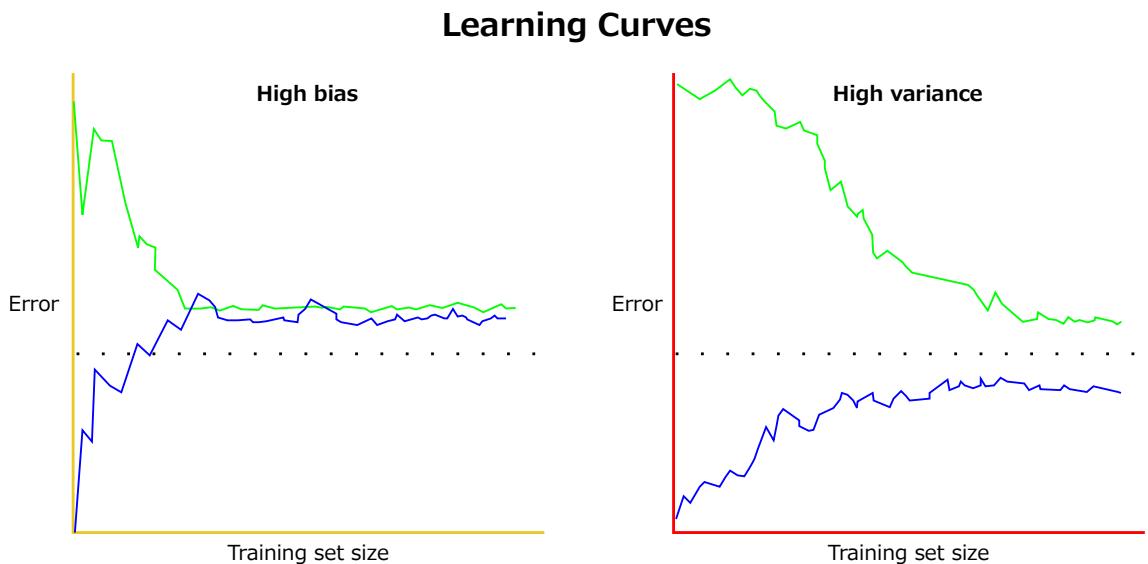
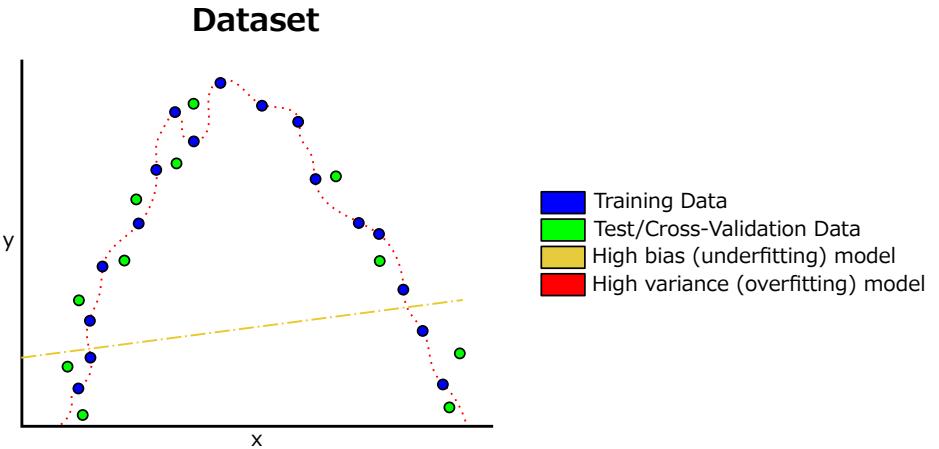


Figure 2.3.1: Plot showing a high bias (underfit) model in yellow and a high variance (overfit) model in yellow. Below are learning curves corresponding to each of these respectively. Learning curves show the effect of different training set sizes on the training and test error of misspecified models. Overfitted models show a large gap between test and training errors, they fit to the training data well but don't generalise to new data (i.e. test data). Underfit models show a very high training error and little difference between test and training data as the model is too simple to fit the training data at all.

unsupervised learning known as semi-supervised learning as well as an alternative idea known as reinforcement learning built on the premise of the psychology of behaviour and the indirect reward of trial and error approaches (Bishop, 2006).

2.3.1 SUPERVISED LEARNING

In supervised (also referred to as predictive) learning the principal aim is to learn a mapping between inputs/features x and outputs/response y from a set of inputs and their corresponding expected output. This is known as the training set i.e. $\mathcal{D} = (x_i, y_i) \quad \forall i \in N$ where N is the cardinality (size) of the training set (Murphy, 2012). A supervised learning algorithm thus seeks to approximate $y = f(x)$ where f is an unknown function. This estimated function $\hat{y} = \hat{f}(x)$ (see 2.3.1) would then generally be applied to new data known as the test data for which the expected outputs are not known (i.e. $x_i \notin \mathcal{D}$).

$$\begin{bmatrix} x_{0,0} & \cdots & x_{0,j} \\ \vdots & \ddots & \vdots \\ x_{i,0} & \cdots & x_{i,j} \end{bmatrix} \xrightarrow{\hat{f}} \begin{bmatrix} y_0 \\ \vdots \\ y_i \end{bmatrix}$$

Supervised learning is further subdivided into two approaches depending on the nature of the expected outputs: classification and regression⁴.

In regression the desired outputs are real-valued (or ordinal) i.e. $y_i \in \mathbb{R}$ and we seek to estimate a particular output quantity for a specific input. The simplest example of this would be the 2-dimensional linear regression problem mentioned above in which we are determining the parameters of a line (gradient/weight and intercept/bias) which best fits the training dataset (\mathcal{D}) composed of pairs of x and y values. Once this line has been found we can use it to predict the value of \hat{y}_i for data in the test set $x_i \notin \mathcal{D}$.

On the other hand, in classification the expected outputs are categorical or nominal variables such as class labels like “host” and “endosymbiont” ($y_i \in \text{host, endosymbiont, ..., C}$). These classifications can be binary (two possible outputs i.e. $y = 0, 1$), multiclass ($|y| > 2$), or multilabel (similar to multiclass but outputs aren’t mutually exclusive, i.e. an input have multiple labels) (Murphy, 2012).

Supervised learning algorithms can also be either probabilistic or non-probabilistic and generative or discriminative. Probabilistic functions will return a probability distribution associated with possible class labels or regression values whereas non-probabilistic approaches will only return the most likely class label or value. Continuing Generative algorithms, such as Naive Bayes or restricted Boltzmann machines, seek to model the process by which the output data was generated from the input i.e. learn the joint probability $p(x, y)$ and make predictions on that basis via Baye’s Theorem (see 2.1).

$$\begin{aligned} p(x, y) &= p(y|x)p(x) \\ p(x, y) &= p(x|y)p(y) \\ p(y|x)p(x) &= p(x|y)p(y) \\ p(y|x) &= \frac{p(x|y)p(y)}{p(x)} \end{aligned} \tag{2.1}$$

Whereas, discriminative classifiers, such as logistic regression/linear classifiers, model the posterior probability $p(y|x)$ ⁵. In other words, for classification problems a generative model would determine the statistical distribution of individual classes whereas discriminative models would just determine the boundaries between them. Generative models often perform better on small training sets by preventing overfitting with discriminative classifiers performing better as the training set grows (Ng and Jordan, 2002).

⁴It is worth noting that the somewhat confusingly named “logistic regression” is typically a form of classification

⁵strictly in cases of bayesian inference this would actually be likelihood as the posterior would be over the parameters i.e. $p(w|Y, X)$ (where Y and X are the whole dataset and w the parameters).

2.3.1.1 SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs) are a type of sparse kernel maximum-margin supervised classification algorithm. With the innovation of the kernel trick in 1992 (Boser et al., 1992) and soft-margins in 1993 (not published until (Cortes and Vapnik, 1995)) SVMs have been among the most successfully applied classification algorithms (Fernández-Delgado et al., 2014). Only relatively recently have they begun to lose ground to the deep learning methods such as deep convolutional neural networks (e.g. LeNet (LeCun et al., 1998)) exemplified by the defeat of SVMs by the LeNet on the MNIST digit recognition dataset (Hinton and Salakhutdinov, 2006; Bengio et al., 2007, 2013).

The goal of SVMs is to learn a hyperplane which separates two sets of labels in the dataset. Note, for multiclass classification a series of one-vs-all classifiers are typically trained (that is for K classes, K SVMs are trained each classifying between a label k and all other labels). However, not all possible hyperplanes that could separate the labels will necessarily generalise well to novel data (and this generalisation is the ultimate goal of supervised learning). Therefore, it is necessary to determine a way to select the hyperplane which should generalise best and to do this in a manner that will be relatively efficient especially with high dimension datasets. This optimal hyperplane for separable classes can be demonstrated to be the hyperplane which maximises the margin between the two classes (Vapnik and Kotz, 1982). In other words, the optimal boundary is the one that has the largest possible distance from each class (while still separating them). Conceptually, the positioning of this boundary is only dependent on the relatively small subset of the training data \mathcal{D} that is near the boundary and it would be inefficient to consider all points when placing the decision boundary. For this reason, SVMs can define the decision boundary in terms of the namesake support vectors and can reformulate their cost function in a more efficient constrained way.

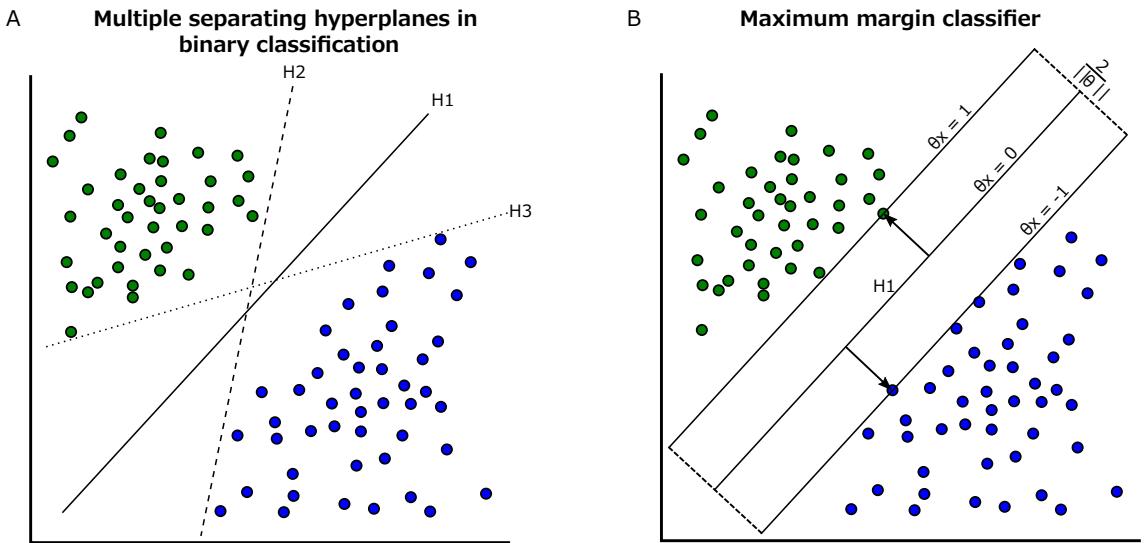


Figure 2.3.2: A: Demonstration of 3 valid decision boundaries in a 2D classification problem, B: The optimal boundary (H_2) is that which maximises the separation of different classes. This optimal boundary can be defined in terms of support vectors. The bias/intercept has been folded into θ directly.

A naive formulation of this problem is simple specifically we are trying to find a linear model $f(x) = \theta_0 + \theta^T x$ which can be simplified to $f(x) = \theta * x$ if we assume that the first element of x is fixed to 1. We thus want to minimise

J in terms of θ to find the largest margin that correctly labels all the training data (in other words is constrained). Fortunately, due to geometry the margin is property of the norm of θ i.e. $||\theta||$ but we use $\frac{1}{2}||\theta||^2$ for mathematical convenience.

$$\operatorname{argmin}_{\theta} \quad J(\theta) = \frac{1}{2} ||\theta||^2 \quad \text{s.t.} \quad y_i(\theta x_i) \geq 1 \quad \forall i$$

In reality, this cost function would be converted to a constrained optimisation problem using Lagrange multipliers and reformulated using the Lagrangian dual form.

The 2nd major enhancement of SVMs is that of soft-margins (Cortes and Vapnik, 1995). Soft-margins are a way of allowing a degree of misclassification if doing so would increase the size of the margin that can be generated. Specifically, a user defined penalty constant C is specified and added to the cost function penalising the degree of misclassification ξ , e.g.:

$$\operatorname{argmin}_{\theta} \quad J(\theta) = \frac{1}{2} ||\theta||^2 + C \sum_{i=1}^n \xi_i \quad \text{s.t.} \quad y_i(\theta x_i) \geq 1 - \xi_i \quad \forall i$$

This can improve robustness to outlier data and generally improve generalisability by keeping the margin as large as possible.

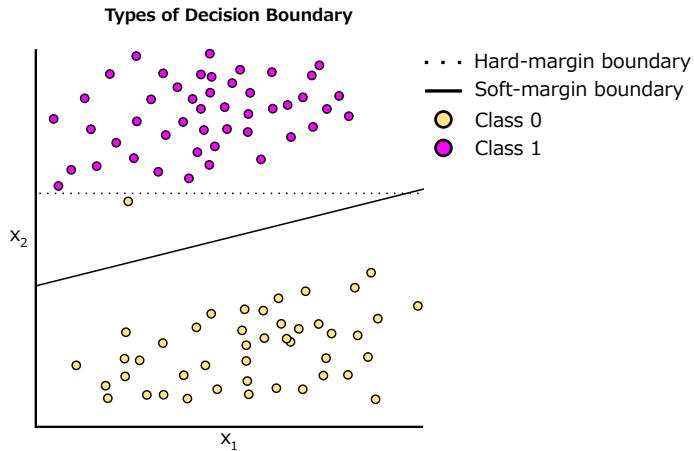


Figure 2.3.3: Demonstration of the utility of a soft-decision boundary to improve the overall fit of a decision boundary by allowing a degree of misclassification during training

Finally, the 3rd major advantage of SVMs is that despite nominally being linear classifiers they can effectively classify data which is not linearly separable in the input dimensions using the kernel trick. Conceptually, a kernel function is used to transform which transforms data from the input dimensions to a higher dimensional space in which the data is linearly separable. These transformed feature spaces can have incredibly high number of dimensions (in the case of popular kernels like radial basis function, an infinite number of dimensions). Explicitly transforming data in this way would be computationally intensive so instead the “kernel trick” is used, where instead of explicitly transforming all the data into the feature space it is done implicitly by computing the inner product of all pairs data points transformed. This is a lot more efficient and precludes the computationally intensive step of converting the data into the new, potentially infinite, co-ordinate space. Radial basis function (RBF) kernel

is an example $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ kernel. Even with the kernel trick, operations on every pair of points can become infeasible for large datasets due to the combinatorial explosion in necessary operations as the dataset increases in size. However, in the same way that the decision boundary parameters are determined using only a subset of the training data (i.e. the support vectors) the kernel trick only needs evaluated on a subset of points near the decision boundary. This is the reason SVMs are sometimes referred to as sparse kernel methods.

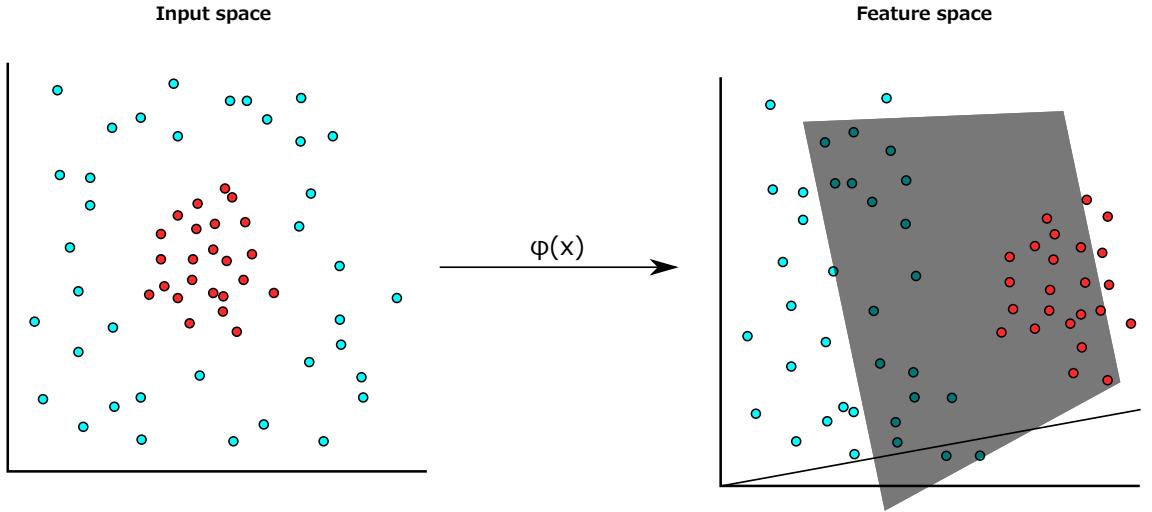


Figure 2.3.4: A kernel transform can allow SVM to produce non-linear classification boundaries by mapping the data to a higher dimensional space in which they are linearly separable. This is known as the kernel trick and the key to its efficiency in SVM is that it is only evaluated for those sets of points near the decision boundary

The advantages of SVM is that they are somewhat resistant to the curse of dimensionality i.e. they are effective with large numbers of features even if the number of features is greater than the size of the training set. By using support vectors, the kernel trick, and Lagrange bound optimisation they are relatively fast and memory efficient to train and as classification only depends on the location of the decision boundary very fast to test. Additionally, in simple form finding the hyperplane of an SVM is a true convex optimisation therefore is guaranteed to always find the global optimum (this guarantee does break with more complex kernels and soft-margins). The major disadvantage is not natively generating probabilistic output (i.e. attaching a probability to a certain classification). However, this can be achieved using methods like Platt Scaling or the related Relevance Vector Machine algorithm. The other disadvantage is that hyper-parameters such as the misclassification penalty for soft-margins (C) and kernel choice (and its parameters) need chosen, typically this is solved by training using a grid-search of permutations of these parameter settings and selecting the best model via cross-validation.

2.3.2 UNSUPERVISED LEARNING

The other main form of machine learning is that of unsupervised or descriptive learning. In which the training dataset has no provided output labels (y) i.e. $\mathcal{D} = x_i \forall i \in N$ (where again N is the cardinality of this training dataset). In other words, we just have our dataset and have no additional information. This is slightly more difficult problem as it lacks an obvious error metric like supervised learning (i.e. difference between actual output and expected output) but is important and useful tool to try to discover patterns in datasets.

There are two major groups of unsupervised learning algorithms, the first of which is clustering algorithms such as K-means that seeks to partition a dataset into a set of groups (see fig. 2.3.5 for more details). The other major group of unsupervised algorithms are those used for visualisation and/or dimensionality reduction. Dimensionality reduction is a way of projecting a multidimensional dataset into a lower number of dimensions in a way that still corresponds to “shape” of the data in the original number of dimensions.

Formally, dimensionality reduction seeks to take a set of data (\mathcal{D}) and convert it to a lower dimension form \mathcal{Y} known as a map $\mathcal{Y} = y_i \forall i \in N$ with each individual x_i in \mathcal{D} represented by a corresponding map point y_i . It also seeks to do this in a way that maintains as much of the structure found in the original data as is possible (Maaten and Hinton, 2008) therefore, if two data points are similar in the original dimensions they should still be similar in the map \mathcal{Y} (and the inverse). Some dimensionality reduction approaches are well known in biology, specifically: principal component analysis (PCA) (Hotelling, 1933) and multidimensional scaling (MDS) (Torgerson, 1952) which both aim to identify hidden features within the dataset that can explain a high degree of the variation.

As ever different methodologies have a range of pros and cons, with some better at preserving global structure (e.g. isomap) and others local data structures (e.g. local linear embedding) and so on. One of the most recent innovations in this area is that of t-distributed stochastic neighbour-embedding (t-SNE) in which the similarity of data points in the input space is modelled as pairwise probabilities using Gaussian distributions. These probabilities are then translated into positions in the map \mathcal{Y} and similarities re-calculated using Student’s t-distributions. The position and variance of these points and distributions respectively is then optimised by minimising the difference between the similarity probabilities in the input space and on the map (Maaten and Hinton, 2008).

2.3.2.1 K-MEANS

K-means clustering is a non-probabilistic unsupervised learning method in which we seek to partition data points in multidimensional space into K clusters. It is often used to initialise Gaussian mixture models.

Specifically, given a set of N observations $X = x_1, \dots, x_N$ of \mathcal{D} dimensions partition each point (x_n) into K clusters

A cluster can be intuitively considered as a group of observations/points which are “closer” to one another than to other observations and the k-th cluster can defined by a \mathcal{D} dimensional vector μ_k , where $k = 1, \dots, K$ for all clusters. This vector represents the current “prototype” centroid of cluster.

So, with k-means clustering we actually seek the set of K cluster centroids μ_k which minimise the sum of squares distances of each data point from its closest cluster centroid. (Bishop, 2006)

If we define a 1-of-K coding scheme with $r_{nk} \in \{0, 1\}$ as a binary variable that is 1 when x_n has been assigned to cluster k (with centroid μ_k) and 0 otherwise then we can define an objective cost function (J) that represents the sum of squares distances of each data point x_n from its assigned cluster centroid μ_k .

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2$$

Therefore, the goal of k-means clustering is to find values for r_{nk} and μ_k that minimise this linear function

2.3.2.1. (Bishop, 2006)

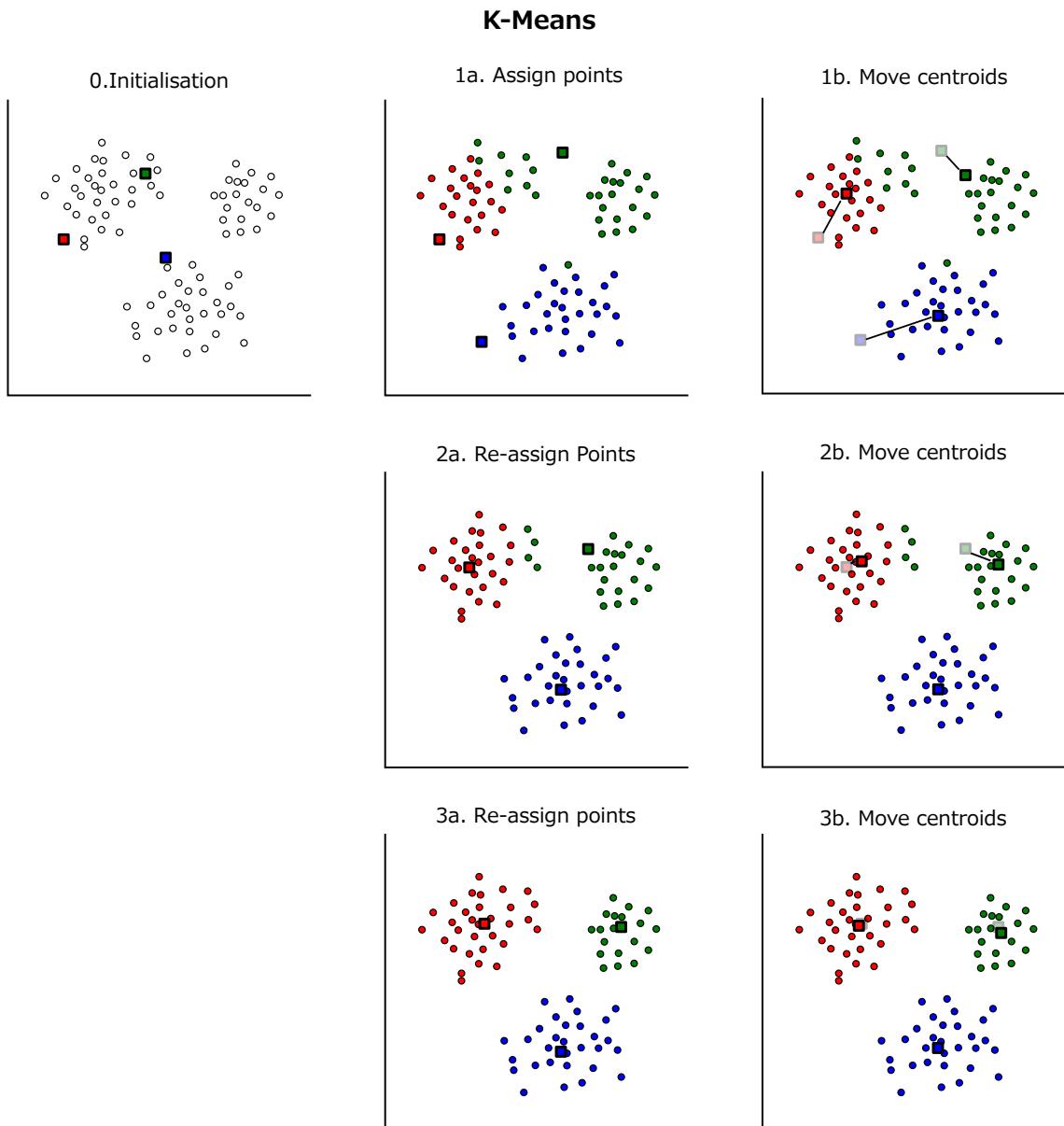


Figure 2.3.5: Demonstration of the K-means algorithm applying 3 rounds of Expectation-Maximisation to cluster a group of 2 dimensional samples. The 3 cluster centroids (represented by coloured rectangles) are randomly initialised in 0 before undergoing 3 rounds of EM. This involves the successive assignment of samples to their nearest centroids (1a, 2a, 3a) and then the movement of the centroids to the center of the points currently assigned to that centroid (1b, 2b, 3b). Assignment of a given sample to a centroid is indicated by a shared colouring and centroid relocation by an arrow with a faded version showing the initial location.

The standard algorithm proceeds in two alternating steps following the initialisation of μ_k with starting cluster centroid locations (Forgy, 1965; Lloyd, 1982):

1. $\operatorname{argmin}_{r_{nk}} J$ i.e. minimise 2.3.2.1 w.r.t the assignment of points to clusters while keeping the cluster centroids fixed.
2. $\operatorname{argmin}_{\mu_k} J$ i.e. minimise 2.3.2.1 w.r.t the position of the cluster centroids while keeping the assignment of points to centroids fixed.

Step 1 roughly corresponds to the expectation step in the expectation-maximisation (EM) algorithm and is trivially achieved by assigning each point to the cluster represented by the nearest centroid or formally:

$$r_{nk} = \begin{cases} 1, & \text{if } k = \operatorname{argmin}_j \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

Step 2 roughly corresponds to the maximisation step in EM is can be determined by taking the partial derivative of J w.r.t μ_k setting it to 0 and solving for μ_k :

$$\begin{aligned} \frac{\partial J}{\partial \mu_k} &= 2 \sum_{n=1}^N r_{nk} (x_n - \mu_k) \\ 0 &= 2 \sum_{n=1}^N r_{nk} (x_n - \mu_k) \\ \mu_k &= \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \end{aligned} \quad (2.2)$$

In other words set μ_k to the mean of all data points x_n assigned to cluster k thus k-means (Bishop, 2006)

These two steps are repeated until a specified maximum number of iterations are reached or no points change cluster assignment during step 1.

J 2.3.2.1 will converge but is liable to get stuck in a local rather than global minimum.

K-means has many modifications and improvements such as refining the initialisation of the clusters by the Bradley-Fayyad method (clustering random samples of the dataset and then k-means clustering the resulting clusters) (Bradley and Bradley, 1998) or over-clustering (running more than k-means clustering with more than the specified number of clusters and merging clusters at the end to generate the correct number of clusters). One of the most recent and promising improvements is that of “ying-yang” k-means clustering which gains a moderate speed-up over the conventional algorithm by minimising the number of distance calculation required. This is achieved by creating upper and lower bound distance filters using the triangle inequality (i.e. $d(a, b) \leq d(a, c) + d(b, c)$ where d is a function that calculates the distance between 2 points) (Ding et al., 2015).

An efficient implementation of the k-means algorithm is available in the MLPACK C++ Machine Learning library (Curtin et al., 2013). While very efficient and effective, k-means has some limitations, it requires a user specified number of clusters and therefore diagnostics to check for obvious misspecification in the number clusters. Information criterion can be used to determine the optimal number of clusters. Additionally, it is not guaranteed to discover the global optimal clusters (can converge to local optima). This can be amortised by running multiple times with different initialisations.

2.4 PHYLOGENETICS

Phylogenetics is an effective tool (if there is sufficient signal/resolution) to investigate the evolutionary ancestry of biological sequence data. It can be used to identify how closely related a given pair of sequences are, as well as indicate what the sequence most likely looked like in a shared common ancestor (ancestral node reconstruc-

tion). Phylogenetic methods also allow estimation of evolutionary processes such as selection pressure, migration, genome reduction, and horizontal gene transfer. In the context of endosymbiosis, phylogenetics can be used to determine evolutionary ancestry of the genes recovered in a transcriptome and to aid identification of the likely origin (host, endosymbiont, contaminant) of these transcripts. Additionally, it can pinpoint potential horizontal gene transfer events between host and endosymbiont by searching for single gene/transcript phylogenies that have an incongruent branching pattern compared to established species trees. Finally, it can be used to aid identification of the putative function of novel transcripts by comparison to other transcripts of known function from databases such as genbank.

Phylogenetics can be defined as a means of arranging a set of character sequences into an optimal hierarchical branching tree structure reflecting some measure of relatedness between the sequences. Usually, these trees will have variable branch lengths that are product of a measure of divergence between the connected nodes.

Typically, these sequences take the form of protein or DNA sequences⁶ and the measure of relatedness is some proxy for evolutionary distance ranging from simple distance measures e.g. Hamming distance ($D = \sum_{k=0}^N |x_k - y_k|$ for two sequences x and y of length N) to more complicated probabilistic estimations based on observed data. A character is an element of a sequence such as an individual base or amino acid, homologous characters are those in separate sequences that are descended from a common ancestor. As they were the first molecular sequences easily available much of the early work in molecular phylogenetics was conducted using protein sequences e.g. ([Eck and Dayhoff, 1966](#); [Fitch and Magoliash, 1967](#)).

This phylogenetic estimation can be a non-trivial process (especially with more complex measures of relatedness) as the number of possible trees rapidly increases with the number of sequences $N_{trees} = \prod_{x=2}^{N_{taxa}} (2x - 3)$. However, the key stages in a phylogenetic analysis are that of sequence sampling (selection of sequences for inclusion in the analysis), alignment (in which homologous sites in the sampled sequences are aligned with one another), masking (in which sites which are evolutionarily informative – can be determined to be homologous but also non-invariant are selected), model selection (in which the best fitting evolutionary model is selected or calculated) and finally, phylogenetic reconstruction (in which the tree is generated that minimises some measure e.g. most likely tree for probabilistic models or least distance).

One implication of most current phylogenetic methods is that they implicitly assume a branching tree structure is the best representative of the evolutionary process that is being modelled. However, as the discovered prevalence of horizontal gene transfer has increased it is becoming that in some cases a network like structure may in fact be more appropriate.

Most analyses in this PhD are conducted using amino acid sequences. DNA is more likely to display a compositional bias, independence of sites is often severely violated due to the structure of codons (3rd base wobble and so on) (non-synonymous mutations more likely to become fixed whereas synonymous mutations are prone to drift). Amino acids also have more states so are less susceptible to back mutations than DNA.

⁶Strictly phylogenetics refers to the study of molecular sequence data although the same methods are applicable to non-molecular characters such as morphological traits (and occasionally originated in this domain) as well as any other set of discrete data vectors. It has even been applied to fields such as linguistics ()

2.4.1 SEQUENCE SAMPLING

Sequence sampling, the selection and identification of sequences for initial inclusion in a phylogenetic analysis, is arguably the most important stage in phylogenetic analysis. Any biases introduced here will propagate throughout the rest of the analysis. While some biases can be mitigated to lesser and greater extents by careful application of various methods in the following stages, there is a degree of fundamental truth in the statement “garbage in - garbage out”.

The aim of proper taxon sampling is to maximise phylogenetic accuracy and to allow testing of specific hypotheses. Phylogenetic accuracy is usually considered in terms of consistency (as data increases the analysis tends towards the correct tree), efficiency (how quickly does this convergence occur), and robustness (how sensitive is the phylogeny to violation of assumptions in reconstruction) ([Nabhan and Sarkar, 2012](#)) Typically, sequence sampling will be conducted from the basis of a single seed sequence which will be used to query existing databases using alignment tools such as BLAST and HMMs (explained below) to attempt to discover potentially homologous sequences from different organisms.

The main issues caused by poor taxonomic sampling in molecular phylogenetics are that of conflicting phylogenetic signals, inadequate rate of evolution to resolve relationships of interest, and violations of assumptions e.g. expectation of a uniform distribution of traits ([Nabhan and Sarkar, 2012](#)).

Generally, increased taxon sampling has a strong positive effect on phylogenetic accuracy ([Zwickl and Hillis, 2002](#)). However, it can also lead to a situation where there are too many sequences to efficiently reconstruct a phylogeny. However, care must also be taken not to unintentionally bias datasets by removing any sequences that are considered “problematic” especially when conflicting phylogenetic signal or model violations can be biologically informative. Therefore, it is usually necessary to include borderline error-generating sequences within a phylogeny initially and to iteratively remove them before repeating the phylogenetic inference. Unfortunately, the reduction of the input sequences to a representative subset by heuristics and/or naive clustering can generate biases of their own. However, tools exist that utilise taxonomic database information to automatically find a set of sequences of a specified cardinality which display the maximum possible taxonomic diversity for that set size ([Zhou et al., 2014](#)).

Another source of bias in sequence sampling is the usually heuristic choice of outgroup taxa. Most contemporary models of phylogenetic inference only infer unrooted trees. Therefore, it is common practice to “root a tree” by selecting a set of sequences from known evolutionarily distant organisms to form an outgroup. If this outgroup is correctly recovered (monophyletically) the root can be placed between it and the other sequences in the phylogeny ([Yang and Rannala, 2012](#)). However, choice of outgroup can change implications which may be drawn from a phylogeny regardless of methodology used to infer it ([Milinkovitch et al., 1996](#)). Care must be taken to ensure the selected outgroup doesn’t actively distort the accuracy of inference of the rest of the phylogeny regardless of the issue of root placement ([Milinkovitch and Lyons-Weiler, 1998](#)).

The two principal ways in which putatively homologous sequences are identified in sequence databases are those based upon Basic Local Alignment Search Tool (BLAST) and its variants and Hidden-Markov Model based

approaches (HMM). Essentially these methods identify and aligning homologous sequences in a target sequence database to our query sequence.

2.4.2 MULTIPLE SEQUENCE ALIGNMENT (MSA)

The goal of MSA is to align sets sequences such that evolutionarily homologous residues occupy the same column. In other words, any given column in the alignment theoretically should contain amino acid or nucleotide residues that derive from the same common ancestor and have evolved in each sequence lineage. It is also possible that insertion or deletion events have taken place and a particular residue is absent in the ancestral node or sequence lineage.

This is a non-trivial computational problem which has been proven to have an NP-complete⁷ computational complexity (Wang and Jiang, 1994). Specifically, the optimal alignment of N sequences has a complexity of $O(L^N)$ for N sequences of length L (Sievers et al., 2011).

Due to this complexity, the majority of MSA algorithms implement heuristic approaches in order to get, if not the optimal solution, a sufficiently good one in a reasonable amount of time.

Typically, MSA algorithms start by generating the sets of all pairwise alignments using established pairwise alignment algorithms. Pairwise alignment algorithms are almost all based upon a pair of “Ur-algorithms” with different goals: Needleman-Wunsch, a global alignment algorithms (which attempts to maximise alignment quality over entire sequence lengths) (Needleman and Wunsch, 1970) and Smith-Waterman, a local alignment algorithm (which is optimised towards producing high quality alignments in sub-strings) (Smith and Waterman, 1981). While early MSA algorithms were typically largely derived from Needleman-Wunsch most modern algorithms seek to combine optimisation of local and global alignments. The distances used in these pairwise alignments will typically be “scored” based upon which matches or alignments are more frequent substitutions (e.g. Leucine and its isomer Isoleucine or Adenine to its fellow purine base Guanine (transition)) are positively scored. Alternatively, gaps (extension of a gap is typically less penalised than creating a gap) or unlikely changes (e.g. the transversion of Adenine to Cytosine or Glutamine to Cysteine) are penalised. This will generally be codified in a substitution matrix e.g. the PAM (Dayhoff et al., 1978), BLOSUM (Henikoff and Henikoff, 1992) amino acid matrices and their numerous subsequent derivations and improvements.

The mostly widely heuristic used to go from these series of pair-wise alignments to a useful MSA is that of progressive-alignment (Feng and Doolittle, 1987) (implemented in tools such as CLUSTAL W (Thompson et al., 1994)). This involves building the pairwise alignment scores into a distance matrix which summarises the relative divergence of each pair of sequences. From this matrix a “guide-tree” is generated using simple neighbour-joining methods (in which a tree is built by recursively clustering the least dissimilar sequences (Saitou and Nei, 1987)). Sequences are then progressively aligned using their branching order within this guide-tree (Thompson et al., 1994).. This drastically reduces the $O(L^N)$ complexity to approximately $O(N^2)$ (Sievers et al., 2011). While there have been various improvements and alternative approaches created such as merging both local and global

⁷A decision problem for which an answer can be verified in polynomial time by a non-deterministic Turing machine and to and from which any NP-hard problem can be translated (Karp, 1972).

alignment (Notredame et al., 2000), rapid identification of homologous regions using Fast Fourier Transforms (Katoh et al., 2002), iterative refinement of alignments (Edgar, 2004b) and use of Hidden-Markov Models (Eddy, 1995).

There have been compelling arguments as early as 1991 that MSA in isolation from phylogenetic inference is inherently flawed as the consideration of evolutionary processes (only really done during phylogenetic inference) is key in the objective weighting and assessment of potential alignments (Thorne et al., 1991). Therefore, the phylogeny and MSA should be jointly inferred (Thorne et al., 1991; Redelings and Suchard, 2005; Bouchard-Côté and Jordan, 2013). This approach also minimises the risk of conscious or subconscious researcher bias towards alignments and subsequent phylogenies that support their pre-conceived ideas. This approach has been attempted using interesting probabilistic programming approaches i.e. BALI-phy (Suchard and Redelings, 2006). Unfortunately, it is still far too slow a process to infer phylogenies in this manner on large or even moderately sized datasets. This means, that for now, independent MSA estimation is here to stay, at least until computational resources and algorithmic development has continued until these more theoretically satisfying approaches become feasible.

Therefore, throughout this thesis, two progressive/iterative alignment tools will be used: Kalign2 (Lassmann and Sonnhammer, 2005; Lassmann et al., 2009) for high-throughput analyses and iteratively refined MAFFT7 (Katoh et al., 2002, 2005; Katoh and Standley, 2013) for individual accuracy critical phylogenetic analyses. Kalign is a very high-speed and relatively accurate (Thompson et al., 2011) progressive alignment tool that uses an efficient and fast Wu-Manber approximate string-matching algorithm to calculate sequence distances (Lassmann and Sonnhammer, 2005). MAFFT, with iterative refinement, is a relatively slow but highly accurate MSA alignment method (Thompson et al., 2011) that incorporates all pairwise alignment information when refining instead of using heuristics to approximate pairwise sequence differences like most approaches.

2.4.3 MASKING

Unfortunately, MSA is far from perfect, especially with the faster algorithms necessary for larger datasets and higher throughput. Therefore, it is often necessary to trim alignments to manually fix any obviously misaligned residues, and remove any ambiguously aligned or absent sites. This has been demonstrated to improve phylogenetic accuracy (Talavera and Castresana, 2007).

However, manual masking can also be a major source of researcher-bias as well as a painstaking process. For this reason, there are tools that attempt to automate this process. They typically score each column independently with criteria including number of absent character states, how similar/variable the character is and if there are multiple putative alignments - how likely is that column to be found in multiple different MSAs. These criteria can then be used to mask out certain columns based on certain thresholds and trade-offs between the length of the alignment and inclusion of low-scoring columns. TrimAL is an example of a tool that automates the masking process using this sort of methodology (Capella-Gutiérrez et al., 2009).

Similarly to MSA, for high-throughput analyses I will use TrimAL whereas for individual accuracy critical analyses masking will be done manually using the graphical tool Seaview (Gouy et al., 2010).

2.4.4 SUBSTITUTION MODEL SELECTION

While the simplest means of phylogenetic inference - parsimony i.e. finding the tree that requires the fewest sequence changes - does not require any explicit model of sequence evolution, all other means of phylogenetic inference do (Le and Gascuel, 2008).

A substitution model is, in its simplest sense, the same as the PAM and BLOSUM matrices used in pairwise and MSA. They are a means of scoring and weighting the significance of different character changes e.g. is an A to a G a more evolutionarily rare state change for a given dataset than an A to a T.

Substitution models typically assume neutrality, independence and finite sites. With the probability of substitution rates having an independent and identically distributed (i.i.d) (Hasegawa et al., 1985). This measure of distance can be naive models where rates of change between character states and the frequency of each state is equal (e.g. $p(x \rightarrow y) \forall x \forall y \in G, C, T, A$ where $x \neq y$ (Jukes and Cantor, 1969)) to models fully parameterised in terms of character frequency and rates of change by the masked alignment (e.g. the generalised time-reversible (GTR) model (Tavare, 1986)). While models like GTR can feasibly be fully parameterised with DNA sequence data due to DNA's relatively few character states it is usually necessary to use empirically-defined models for amino acid datasets. These are substitution matrices that have been determined using the empirically observed substitution rates for various amino acids changes in many large MSAs (Le and Gascuel, 2008).

Unfortunately, a single substitution model will rarely hold true over an entire alignment with the rate of evolution varying both across and within sites (heterogeneity and heterotachy). The frequency of character states also frequently changes across a phylogeny. It is important to control for these phenomena, because, as mentioned earlier, violation of model assumptions can decrease phylogenetic accuracy.

The most frequent violation that is controlled for is allowing the rate of substitution to vary across sites by using a Γ distribution $Var = \frac{\alpha}{\beta^2}$, $\mu = \frac{\alpha}{\beta}$ with a given shape α and trivial scale factor β depending on the dataset to scale rates at each site. For datasets that have a high degree of rate heterogeneity a low valued α produces a broad distribution of rates, whereas a high value will generate a narrow distribution for datasets with low rate heterogeneity (Yang, 1993). For reasons of computational efficiency Γ is typically approximated as a discrete distribution of 4 to 8 categories of equal probability (Yang, 1994). A more limited version of this is the invariant sites model in which sites are divided into 2 classes, one considered invariable while the other has normal substitution rates applied⁸(Hasegawa et al., 1985).

Unfortunately, these models still assume other model parameters namely the equilibrium frequencies and relative rates are the same across sites (but just scaled). However, some models have been proposed with multiple rate matrices (Lartillot and Philippe, 2004) and state frequencies can be defined at each site (Bruno, 1996) but needs lots of taxa (Lartillot and Philippe, 2004). An alternative to this is the CAT model which a mixture model mixture model with K classes each containing a different state frequency. If $K = N$ then this is the same as Bruno's model however, generally $K < N$. A probabilistic process known as a Dirichlet Process Prior is used to assign columns to various state frequency classes and simultaneously determines the optimal value of K during

⁸This can also be used with Γ and is approximately equivalent to the addition of another discrete Γ category.

this process (Lartillot and Philippe, 2004). An alternative to this approach is explicitly partitioning a masked alignment and generating a model and state frequencies for each partition, some consider this equivalent to a CAT model depending akin to preferences for fixed-effects vs random-effects models (Yang and Rannala, 2012). However, personally, automated partitioning using a Dirichlet process has the advantage of not requiring arbitrary user-defined partitions, which could be a source of bias.

Finally, the rate of evolution can vary even with a site itself (a process known as heterotachy) especially when large numbers of divergent taxa are included in a masked alignment. One model modification which attempts to control for this is that of the covarion model. It allows sites to switch between on and off using an infinite mixture model. The proportion of on and off sites is determined at each site (Zhou et al., 2010).

Generally, simpler models such as the “null” parsimony model or basic models that don’t account for complex evolutionary phenomena are more susceptible to artefacts such as long-branch attraction (LBA)⁹ (Yang, 1996).

However, in the grand tradition of “no-free lunch”, there is no universally best model for all datasets. Therefore, it is necessary to test multiple competing models using a provided MSA. Typically, these models are then compared for their fit to the observed data using information criterion (Sullivan and Joyce, 2005) such as Akaike’s (AIC) which assess fit while penalising model complexity in a standard regularisation trade-off ($AIC = 2k - 2\ln(L)$ where k is the number of parameters and L the model likelihood (Akaike, 1974)). Other criteria include corrected AIC (Sugiura, 1978), Bayesian Information Criteria (Schwarz, 1978) and Decision Theoretic criteria (Minin et al., 2003) based approaches (Sullivan and Joyce, 2005).

Throughout this thesis, I will use 2 tools which incorporate these various criteria to infer the best fitting model depending on the input data. ProtTest3 (Abascal et al., 2005; Darriba et al., 2011) will used for analyses involving protein sequences and jModelTest2 (Posada, 2008; Darriba et al., 2012) for phylogenetic inference of DNA datasets.

2.4.5 PHYLOGENETIC INFERENCE

The simplest phylogenetic inferences are those of distance matrix methods. Distance matrix methods (Fitch and Magoliash, 1967) work on the basis of generating a matrix representing the pairwise distances of each sequence using the selected substitution model and inferring a phylogeny from this. The simplest case would be searching tree space for the optimal tree using a standard least-squares criteria between actual and expected branch lengths (i.e. distances) (Fitch and Magoliash, 1967; Cavalli-Sforza and Edwards, 1967).

However, the most common is that of neighbour-joining which begins with the distance matrix and a star topology tree in which all leaf node branches are connected to a single shared central node. Then:

1. Find the closest two branches in the distance matrix
2. Join the closest pair into a single branch with a new internal node connected to central node

⁹LBA is a distorting effect in which long branches (rapidly diverging) are incorrectly placed close to one another regardless of actual shared homology. This is due to the increased chance of rapidly diverging sequences to share independently acquired residues (Bergsten, 2005).

3. Generate a new distance matrix reducing the selected pair to the new node (the distance of the selected pair of leafs to the new internal node and the distance between every other leaf and the new internal node)
4. Repeat the process with the new matrix (Nei, 1987)

It works on the assumption that the true tree has the smallest expected length (minimum evolution) and a short tree that has similar topology can be achieved using the fast simple agglomerative algorithm. NJ is one of the best distance methods and is more reliable than maximum-parsimony which can be asymptotically inconsistent. While already efficient (possibly efficient as possible) NJ can be made more efficient using effective heuristics to search tree space (Kumar, 1996). As well as improvements where variance is minimised instead of pure distance improving performance in datasets with high substitution rates e.g. BIONJ (Gascuel, 1997).

Distance methods are very fast but can perform very poorly for divergent sequences with large sampling errors as they don't generally account for variance in distance estimates (Yang and Rannala, 2012) (BIONJ partially adds this). They are also particularly sensitive to gaps in the alignment.

Parsimony approaches (Camin and Sokal, 1965) on molecular sequences (Eck and Dayhoff, 1966) seek to infer the maximum parsimony (MP) tree. That is the tree which requires the smallest number of character changes (has the best tree score). Where the tree score is the sum of all character lengths (the minimum number of changes for each site in the alignment). Any site that is invariable is not informative for generation of a parsimony tree. It has no explicit assumptions relative to other methods however, this means it is difficult to build in prior knowledge of sequence evolution when generating a tree. It also fails when multiple substitutions have occurred at the same site or with parallel changes in two long branches and therefore is especially prone to long-branch attraction (Felsenstein, 1978). Prestige of parsimony methods declined with discoveries that they can produce statistically inconsistent phylogenies (Felsenstein, 2001).

A majority consensus tree will typically be presented with each node annotated with the number of bootstrapped trees that supported its existence.

2.4.5.1 MAXIMUM LIKELIHOOD

Maximum likelihood (ML) methods seek to discover the maximum-likelihood estimates (MLEs) of the tree parameters (topology τ , branch length θ_l and usually substitution model parameters θ_μ) for the data i.e. MLE of $L(\tau, \theta_l, \theta_\mu)$.

These MLEs are estimated numerically using standard iterative optimisation algorithms. They were developed relatively early in molecular phylogenetics using relatively simple models (Neyman, 1971) but more efficient implementations, e.g. (Felsenstein, 1981), and increased computational power has made them one of the more popular means for phylogenetic inference.

Generally, an ML approach will sequentially perturb a starting tree topology (often BIONJ or simple ML tree itself) using branch swapping operations such as Nearest-Neighbour Interchanges (NNI) or Subtree-Prune-and-Regraph (SPR) where whole subtrees are removed and reattached to a different part of tree. SPR is slower but less prone to get caught in local optima than NNI and thus will lead to higher likelihood phylogenies overall (Criscuolo,

2011). Expectation-maximisation can then be used to find the MLE for branch length and model parameters. For example, PhyML uses an initial BIONJ and standard hill-climbing which perturbs topology and branch lengths simultaneously

The advantage of ML approaches is that they have explicit model assumptions (which can therefore be tested), are relatively robust to model misspecification, are relatively efficient in a phylogenetic sense, and can make use of sophisticated evolutionary models. They can thus compensate for many pathological data features (heterotachy, state and rate heterogeneity within and across sites) (Yang and Rannala, 2012). Almost all published phylogenies are Bayesian or ML (or ideally both) for this reason. Unfortunately, ML inference is also relatively slow to calculate especially in comparison to distance methods.

In order to get an estimate for the robustness of a particular phylogenetic ML inference, the masked alignment can be repeatedly resampled (bootstrap samples) with replacement and phylogenies regenerated. Each node can then be scored based on the proportion of these bootstrap samples in which it is recapitulated (Felsenstein, 1985). A similar approach, known as jackknifing, uses random subsets of the alignment instead of samples (Miller, 1974; Lapointe et al., 1994). Finally, approximate likelihood-ratio tests (aLRT) can be used on branches to give support values by comparing the likelihood of existence of a given branch compared to its non-existence (Anisimova and Gascuel, 2006). There is considerable literature evaluating the pros and cons of different support schemes. However, bootstrap supports are the *de facto* method of inferring a variance of phylogenetic error (Stamatakis et al., 2008) as they are both simple and conservative (but are computationally expensive) (Anisimova and Gascuel, 2006). It should be noted that all of the above methods of determining phylogenetic robustness can be applied to distance and parsimony methods as well.

In this work, for high-throughput analyses FastTree2 was due to its considerable greater speed relative to ML inference tools (Price et al., 2010). For individual phylogenetic analyses ML trees were inferred using RAxML8 (Stamatakis, 2014) and non-rapid bootstrap supports.

2.4.5.2 BAYESIAN

Bayesian inference is, as the name suggests, based on the Baye's theorem. $p(\tau, \theta_l, \theta_\mu | D) = \frac{p(\tau, \theta_l, \theta_\mu)p(D|\tau, \theta_l, \theta_\mu)}{p(D)}$ with $p(\tau, \theta_l, \theta_\mu)$ being the prior probability for model parameters (topology τ , branch length θ_l , substitution model θ_μ) $p(D|\tau, \theta_l, \theta_\mu)$ being the likelihood of the data given a certain set of parameters with $p(D)$ as the marginal probability.

Due to the computational difficulty directly calculating the marginal likelihood (integrated over all possible parameter values in all dimensions) phylogenetic inference uses a process known as Monte-Carlo Markov-Chains (MCMC) to sequentially randomly sample the posterior probability distribution. Conceptually, these can be considered as random walkers on the probability distribution that are more likely to accept new movements that increase likelihood than decrease.

An advantage of Bayesian inference is that the posterior probability (PP) of a given node means "support" values are built-in to the inference and additional bootstrapping is unnecessary. Unfortunately, posterior probabilities are sensitive to model violations and have been found to not be very conservative estimators (Simmons,

2003) (although again there has been considerable work comparing Bootstraps to PP (Anisimova and Gascuel, 2006)). Additionally, the prior distribution in Bayesian Inference allows information that is already known about the dataset to effectively be built-in to the inference. This potentially improves phylogenetic accuracy.

All in-depth individual phylogenetic analyses presented in this thesis were inferred using both Bayesian (via MrBayes3 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012)) and ML methods. Phylogenies are then presented with both PP and bootstrap support values on each node inferred using both methodologies.

2.5 INFORMATICS LANGUAGES AND HARDWARE

2.5.1 LANGUAGES AND LIBRARIES

Several programming languages and a range of libraries were used throughout this PhD depending on the suitability of a particular tool for a task. The full details of the specific tools used for the main analyses are outlined during the description of these analyses, however, the tools used for prototyping as well as those used for smaller tasks not covered in detail are omitted elsewhere.

Languages and libraries were chosen depending on their best fit for a particular task. Performance sensitive code such as those dealing with large datasets (e.g. high-throughput sequencing libraries or image data) were principally conducted using the C++ language in line with the C++11 standard (ISO International Standard, 2011)

The main C++ libraries used in addition to the C++11 standard library were:

- Seqtk - fastq/a sequence parsing library (Li, 2015)
- MLPACK - a high-performance machine learning library <http://www.mlpack.org/> (Curtin et al., 2013)
- OpenCV3 - widely used computer vision library (Bradski, 2000)
- Armadillo - numerical computation library (Sanderson, 2010)

The majority of tasks were accomplished using the high-level python language (python2.7 or python3.4 depending on the application). In addition to the standard library, the numerical computation libraries numpy and theano, machine learning library scikit-learn, statistical and scientific libraries scipy and pandas, the bioinformatics libraries scikit-bio and biopython, and plotting libraries holoviews, matplotlib, seaborn and bokeh were all used extensively. Frequent use was made of literate programming offered by environments such as the ipython notebook (recently renamed jupyter).

The statistical programming language R (R Core Team, 2015) was used for some data analysis and visualisation primarily using ggplot2 (Wickham, 2009) and dplyr (Wickham and Francois, 2014). This was primarily done using the R-Studio <http://www.rstudio.com/> integrated development environment) and R-markdown (Alaire et al., 2014).

All code was version controlled using git <http://git-scm.com/> and remotely hosted using github <https://github.com/> and bitbucket <https://bitbucket.org/> services. Unit tests were automatically run on syn-

chronisation ('push') with these remote servers using the Travis <https://travis-ci.org/> Continuous Integration service.

Incidental scripting was done using zsh and bash languages and all code was written using a simple vim terminal.

2.5.2 HARDWARE

All analyses were conducted on either the lab cluster (running Ubuntu Server LTS 12.04 and 14.04 <http://www.ubuntu.com>:

- PowerEdgeM910 with 2 x Intel Xeon CPU E6510@1.73GHz, 512GB RAM
- PowerEdgeM910 with 2 x Intel Xeon CPU E7 – 4807@1.87GHz, 512GB RAM
- PowerEdgeM620 with 2 x Intel Xeon CPU E5 – 2650v2@2.60GHz, 512GB RAM

Or on two workstations both running continuously updated versions of Arch Linux <https://www.archlinux.org/>.

- Apple MacPro with 2 x Intel Xeon CPU E5520@2.27GHz, 16GB RAM
- Dell Precision T7500 with 2 x Intel Xeon E5620@2.4GHz, 48GB RAM

*Taxonomy is described sometimes as a science and sometimes as an art,
but really it's a battleground*

- Bill Bryson: *A Short History of Nearly Everything*

3

Endosymbiont diversity

3.1 INTRODUCTION

3.1.1 ENDOSYMBIONT TAXONOMY AND CLONALITY

Over 50 strains of green algal photobionts have been identified in *Paramecium bursaria* species (Hoshina et al., 2010, 2004; Hoshina and Imamura, 2009; Summerer et al., 2008; Vorobyev et al., 2009). These form at least 4 distinct species groups, believed to be represented in the following cultures:

- *Micractinium reisseri* (e.g. former “European” group endosymbionts such as those attributed to CCAP 1660/12)
- *Chlorella variabilis* (e.g. former “American” group endosymbionts such as *Chlorella variabilis* NC64A)
- *Chlorella vulgaris* (e.g. the endosymbiont attributed to CCAP 1660/10)
- *Coccomyxa* sp. (e.g. the endosymbiont attributed to CCAP 1660/13)

These species display a polyphyletic distribution within the green algae providing evidence for multiple separate origin events for the *P. bursaria* endosymbiosis (Hoshina and Imamura, 2008, 2009) Furthermore, there is emerging evidence, in the form of intron HGTs and ITS2 sequencing data that strains of *P. bursaria* are capable of hosting double and triple co-habitations of different photobiont species (Hoshina, 2012). Therefore, before an

effective analysis can take place of an endosymbiotic system it is important to carefully define the species (singular or plural) involved.

Unfortunately, the systematics of the Chlorophyta has experienced a relatively high degree of flux, with multiple redefinitions even since the initial use of molecular phylogenetics of ribosomal sequences (Hori et al., 1985; Gunderson et al., 1987) in the 1980s (Leliaert et al., 2012; Hoshina et al., 2010). The algal endosymbionts of *Paramecium bursaria* in particular have gone through a range of names and classifications starting with *Zoothiorella* in 1882 and through various species of the genus *Chlorella* (Hoshina et al., 2010).

Initially, all symbiotic algae were named as single *Chlorella parameci* species but this name was rejected and *Chlorella variabilis* was defined (Shihira and Krauss, 1965) but this was in turn rejected and fell out of use. Later, the first discovery of the existence of multiple distinct strains of photobiont was published (Douglas, 1986). With this came the understanding that the endosymbionts of *P. bursaria* are likely to be divergent but not distinct species to other described free-living *Chlorella* (Hoshina et al., 2010).

To add further confusion to the system, the most recently accepted terms defined species of endosymbiont merely as “American” and “European”. This lead to several misidentifications (Kodama et al., 2007; Hoshina et al., 2010). Recently, these two organisms have been redescribed as distinct species *Chlorella variabilis* and *Micractinium reisseri* respectively (Hoshina et al., 2010). Therefore, care must be taken when reading older literature to distinguish the earlier less well-defined *C. variabilis* from the modern usage.

One source of complication in the systematics of the photobionts have been plagued with cases of mislabelling and loss of cultures by various culture collections and in papers. For example, the initial culture which the original *Chlorella variabilis* was described from was lost and a supposedly identical culture from a different collection was found to have wildly different biochemical properties (Hoshina et al., 2010). These complications and confusions add to the importance of accurate endosymbiont species identification.

3.1.2 ITS2 TAXONOMIC PROFILING

The most widely accepted means of rapidly taxonomically profiling archaeplastida (and indeed a range of eukaryote species) is that of nuclear ribosomal internal transcribed spacer 2 (ITS2) (see fig. 3.1.1) barcoding. ITS2 has shown particular utility in the identification and separation of closely related green algal species (Buchheim et al., 2011; Heeg and Wolf, 2015) due to being universal, reliably amplifiable and highly variable (Hershkovitz and Lewis, 1996).

ITS2 barcoding has been recommended as a superior marker to other universal archaeplastida DNA barcodes such as the *rbcL* (Chen et al., 2010). The conserved nature of the flanking 5.8S and 18S sequences allows near universal primers to be designed which efficiently amplify ITS2 sequences unlike the broadly distributed but highly variable *rbcL* (Buchheim et al., 2011).

In many species the rDNA cistron is present in multiple copies as tandem head-to-tail repeats varying in copy number from 1,2 copies to thousands (Torres-Machorro et al., 2010). While these copies are frequently homogeneous there are many organisms that display intranuclear variation (Buchheim et al., 2011). For example, alveo-

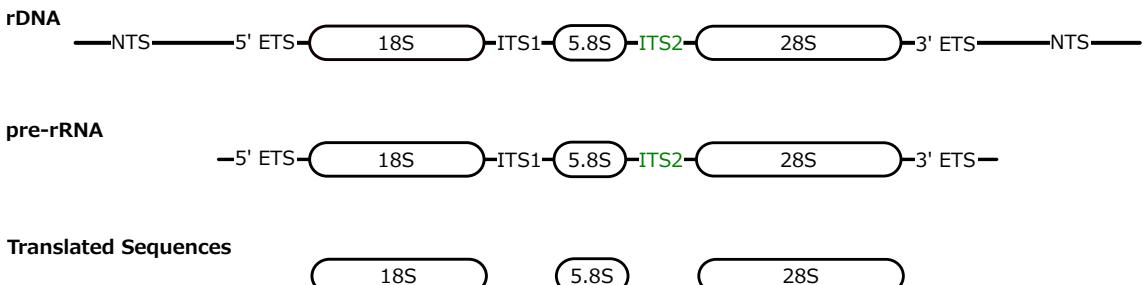


Figure 3.1.1: Structure of Eukaryotic nuclear ribosomal DNA. rRNA genes exist in tandem repeats separated by nontranscribed spacers (NTS). These NTS are composed of internally transcribed spacers (ITS) and externally transcribed spacers (ETS). ITS2 is highlighted in green and forms an effective taxonomic barcode at sequence level for eukaryotic species analyses. The ITS2 secondary structure shows a greater level of conservation and can be used to investigate lower distance systematic relationships. Figure was redrawn from ([Shi, 2005](#))

lates have been discovered with variant rDNA copies ([Stern et al., 2012; Galluzzi et al., 2004](#)). At different points in the life cycle of *Plasmodium* spp. ([Nishimoto et al., 2008](#)) there is expression of SSU rRNA gene paralogues with up to 11% difference ([McCutchan et al., 1988; Chambouvet et al., 2015](#)). Similarly, chlorophytes have previously displayed heterogeneity in rDNA copies ([Pillmann et al., 1997; Fama et al., 2000](#)). Therefore, care must be taken not to assume intranuclear homogeneity in phylogenetic analysis of ITS2 sequences ([Buchheim et al., 2011](#))

This said, ITS2 sequencing will identify the endosymbiont species in the CCAP 1660/12, Yad1g1N and CCAP 1660/13 cultures. It will also offer a tool in which to investigate the presence of clonality within the photobiont populations. By amplifying and sequencing a large number of ITS2 fragments from the same culture there is a reasonably good chance that all the ITS2 level diversity will be sampled. If on analysis these sequences form multiple clades or display divergent groupings this could be strong evidence for a multiple photobiont co-habitation within the *P. bursaria* host.

Finally, one last means in which we sought gain additional insight into the host-endosymbiont system was through the use of multiple-displacement amplification (MDA) based sequencing ([Lasken, 2007](#)). Due to difficulties in obtaining sufficient culture densities and the prevalence of putative sources of contamination within the culture bulk genome sequencing was considered to be prone to major difficulties. Therefore, MDA offered a way in which we could further investigate this question of photobiont clonality while also generating a resources with potential use for further analysis. For example, searching for potentially biologically significant genes that are present in the genome but are not transcribed during endosymbiosis. The utility of this genomic resources hinges on our ability to partition recovered genomes/contigs into the originating host and endosymbiont genomes. It is particularly important to do this and effectively discard contaminant contigs derived from bacteria (food and symbionts) and viruses associated with the host.

3.1.3 ISOLATION OF *P. BURSARIA*

One avenue that is important for an effective analysis of a host-endosymbiont system is the ability to analyse the partners in isolation. This can be used to test individual hypotheses regarding each partner and allows controlled

reintroduction experiments to be undertaken. Unfortunately, the majority of extant, well-characterised endosymbioses display metabolic co-dependence and therefore, host and endosymbiont cannot be isolated without one or other dying (i.e. they form an obligate relationship as supposed to a facultative one).

Fortunately, there have been numerous studies that have investigated the separation of host and symbiont in *P. bursaria* - green algal systems e.g. (Hosoya et al., 1995; Achilles-Day and Day, 2013b; Karakashian, 1963). Most recently, the only transcriptomic analysis of this system by (Kodama and Fujishima, 2014). investigated the differential global metatranscriptome profile of *P. bursaria* Yad1g strain with and without its *Chlorella variabilis* 1N endosymbiont (Kodama and Fujishima, 2014). While, this is a different strain of both host and endosymbiont to the SW1-ZK strains in the CCAP1660/12 culture (*P. bursaria* and *Micractinium reisseri*) reproduction of this endosymbiont clearing offers a potential avenue by which to further investigate and, combined with RNAi to test the functional underpinning of this relationship.

There have been several published methods for clearing endosymbionts from host cells namely, the herbicide paraquat (Hosoya et al., 1995), culturing under constant dark (Karakashian, 1963), herbicide DCMU (Kodama and Fujishima, 2009), X-ray (Wichterman, 1948), and cycloheximide (Weis, 1984; Kodama et al., 2007). Therefore, we attempted 3 of these methods: specifically Paraquat, Cycloheximide, and constant darkness treatments with bacterial feeding in order to clear endosymbionts from the host *Paramecium*.

3.2 AIMS

In this chapter I will determine the exact algal endosymbiont strains present in the principal *Paramecium bursaria* cultures used throughout this thesis and their relationships relative to one another and to other green algae.

I will also use this data and single cell genomics to investigate whether the algal endosymbiont present in the *Paramecium bursaria-Micractinium reisseri* CCAP 1660/12 strains form a clonal population.

Finally, I will discuss the attempts to remove the endosymbiont in the *Paramecium bursaria* CCAP 1660/12 strain from the host.

3.3 METHODS

3.3.1 TAXONOMIC ANALYSIS

3.3.1.1 ITS₂ SEQUENCING

Paramecium bursaria CCAP 1660/12 and *Paramecium bursaria* CCAP 1660/13 cultures were maintained in New Cereal Life (NCL) media at 18°C with 12:12 hour light/dark cycle. In order to mitigate the risk of sequencing free-living algae in the CCAP 1660/13 culture, ITS₂ sequences were acquired from both pure culture samples and carefully purified samples. Purification involved successive filtering and washing steps of isolated cells in sterile NCL media. Specifically, filtration using a 10µm filter, washing off, re suspension and 3 serial subcultures in sterile NCL media.

ITS₂ sequences were amplified using 2 pairs of primers: ITS₂-S₂F primer (ATGCGATACTGGTGTGAAT)

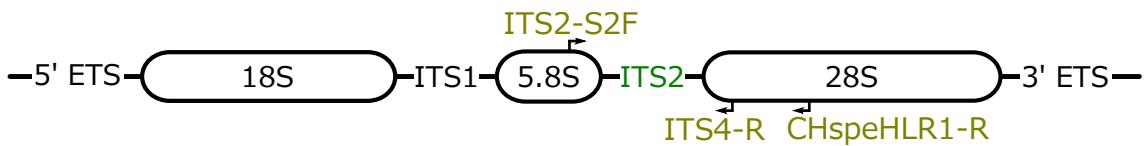


Figure 3.3.1: Schematic diagram showing the location of the forward (ITS2-S2F) and reverse primers (CHspeHLR1R, ITS4-R) used for the amplification of ITS2 sequences in this study. CHspeHLR1R binds within the 28S whereas ITS4-R binds closer to the 5' end of the 28S. Both primer sets recover the full ITS2 sequence.

binding to conserved 5.8S sequences from (Chen et al., 2010) with the CHspeHLR1R (CACTAGACTACAATTGCCAGGCC) reverse primer specific to chlorophyte 28S (Hoshina et al., 2004) and the ITS4 primer

(TCCTCCGCTTATTGATATGC) (White et al., 1990) (see fig. 3.3.1). The reason for the dual primer approach was that it was observed in the smaller biological samples created during the cleaning process that the ITS2-S2F - CHspeHLR1R primer pair wasn't amplifying ITS2 very efficiently therefore the alternate primer pair was used.

PCR conditions used were 94°C for 5 minutes followed by 40 cycles of 30 seconds at 94°C, 30 seconds at 56°C, and 45 seconds at 72°C. This was followed by a final elongation step of 10 minutes at 72°C.

PCR products were then cleaned up, cloned, sequenced and processed using the same protocol as (Maguire et al., 2014). Briefly, the successfully amplified PCR products were gel-purified (Wizard SV Gel and PCR Clean-Up kit, Promega). These products were then TA-cloned using Agilent's PCR StrataClone Cl e-white screened and 5 clones selected for each PCR product. Clones were then externally Sanger sequenced using the M13Rev primer at MWG Eurofins. Flanking vector and primer sequences were removed: sequenced trimmed to areas of high chromatograph quality and ambiguously defined bases corrected using Sequencher (Gene Codes, 2015).

From the 3 *Paramecium bursaria* CCAP 1660/12 biological replicates 14, 9, and 11 ITS2 sequences were obtained respectively. Similarly, from the 2 *Paramecium bursaria* CCAP 1660/13 biological replicates 8 were obtained from sequences obtained from the culture directly, and 10 from the purified, washed samples (7 using ITS2-S2F-ITS4 primers and 3 using ITS2-CHspe). Additionally, 5 ITS2 sequences were acquired from the Yad1g1N culture following the same protocol.

In order to mitigate the risk of sequencing error masquerading as true sequence divergence any sequences found in later phylogenetic analysis to demonstrate single nucleotide changes from the consensus of its clade placement was resequenced at MWG Eurofins in reverse using M13Uni. Specifically, these were ITS-B18, ITS-2, ITS-19, ITS-B6, ITS-B3, ITS-A7, ITS-6, ITS-B15, ITS-10, ITS-9, ITS-15, and ITS-1.

3.3.1.2 PHYLOGENETICS

ITS2 sequences used in (Hoshina et al., 2010), (Hoshina and Fujiwara, 2013) were retrieved from genbank. The trimmed sequences and the established database sequences were then aligned using MUSCLE (Edgar, 2004a). This alignment was manually masked in the graphical SeaView (Gouy et al., 2010) package. jModelTest2 (Guindon and Gascuel, 2003; Darriba et al., 2012) was then used to pick an appropriate substitution model. Finally, phylogenies were inferred using the maximum likelihood method via RAxML version 8 (Stamatakis, 2014) with

1,000 bootstrap replicates. Similarly, MrBayes (Huelsenbeck and Ronquist, 2001) was used to infer the phylogeny using the bayesian formulation. MrBayes used 2 independent runs of 4 Monte-Carlo Markov-Chains (MCMC) for 3,750,000 million generations (at which point the 2 runs were considered to have converged, as determined in Tracer v1.4 (Rambaut and Drummond, 2007)). Trees were estimated from the MCMC results with a burn-in of 250,000 generations. Trees were then visualised and support values combined using TreeGraph2 (Stöver and Müller, 2010).

Trees were rooted using a Microthamiales outgroup composed of *Trebouxia gigantea* AJ249577.2, *Trebouxia arboricola* SAG219-1a Z68705.1, *Trebouxia jamesii* Hp-MT1 AJ511357.1, *Trebouxia impressa* AJ249576.1, *Trebouxia corticola* AJ249566.1, and *Trebouxia higginsiae* AJ249574.1.

3.3.2 SINGLE CELL GENOMICS

3.3.2.1 DNA EXTRACTION

Individual *P. bursaria* CCAP 1660/12 cells were removed from culture and washed three times in a successive series of 10 μ l drops of sterile modified New Cereal Leaf-Prescott (NCL) medium to minimise prokaryotic contamination from bacterial foodstocks in the culture media. Cells were added to a final 10 μ l drop of sterile water before being added to a microcentrifuge tube.

DNA was then extracted using a Cetrimonium bromide based method adapted from (Winnepenninckx et al., 1993). In brief, 748.5 μ l of CTAB extraction buffer (at 37°C and 100 μ l beads (Sigma, 425600 μ m; acid washed) were added and the tube was vortexed for 5 minutes. The tube was incubated for 50 minutes at 37°C, vortexed again for 5 minutes and incubated for 50 minutes at 60°C. This was to ensure lysis of the endosymbiont chitin cell wall. DNA was extracted three times with phenol/chloroform/isoamylacohol (25:24:1, pH 8), washed with 70% ethanol and re-suspended in 2.5 μ l TE (pH 8). Whole-genome amplification of purified genomic DNA was performed using the multiple-displacement amplification based (MDA) Qiagen REPLI-g Single Cell Kit. The REPLI-g amplified gDNA was purified using a QIAamp DNA mini kit and eluted in 100 μ l elution buffer.

3.3.2.2 SEQUENCING

5 prepared libraries were put forward for sequencing (Pb-3, Pb-4, Pb-6, Pb-7 and Pb-8). Samples were multiplexed and were rapid sequenced in an Illumina HiSeq 2500 in 150bp paired-end mode.

3.3.2.3 READ PRE-PROCESSING

Trimmomatic (Bolger et al., 2014) was used to trim sequencing adapters (using sequences provided by Exeter Sequencing Service) via the ILLUMINACLIP setting. Reads were then quality trimmed at a minimum average SLIDINGWINDOW quality thresholds of Q5 and Q30.

Q5 and Q30 trimmed reads were then error corrected using BayesHammer (Nikolenko et al., 2013) as built into the SPAdes assembler (Bankevich et al., 2012).

Trimmed and error corrected libraries were also then digitally normalised (Brown et al., 2012) to a coverage of

20 and with a K-mer size of 25. K-mers were then abundance filtered (Zhang et al., 2014, 2015) using the Khmer package (Crusoe et al., 2015).

3.3.2.4 ASSEMBLY

Assemblies were then generated using the following sets of data:

- Q₅ trimmed reads with error correction
- Q₃₀ trimmed reads with error correction

The following assemblers were used:

- SPAdes assembler (Bankevich et al., 2012; Nurk et al., 2013)
- SPAdes assembler with “careful” thresholding (runs MismatchCorrector and minimises the risk of indels)
- MEGAHIT (Li et al., 2015)
- Platanus (Kajitani et al., 2014)

3.3.2.5 ASSEMBLY ASSESSMENT

Assemblies were assessed and compared using the QUality ASsessment Tool for genome assembly (QUAST) (Gurevich et al., 2013) and key assembly metrics were compared (N₅₀, N₉₀, contig number and length and total assembly size).

3.3.2.6 CONTIG BINNING

Contigs were subsequently cut into 10kb fragments for consistency in binning and taxonomic assignment and obviate the difficulties aligning very long sequences. Reads were then mapped back onto the final assembly using Bowtie2 (Langmead and Salzberg, 2012)

Using the metagenomic binning tool, CONCOCT (Alneberg et al., 2014) contigs were binned into clusters based on sequence composition and coverage features (derived from mapping data). Coverage features were derived from a coverage and linkage table generated via CONCOCT scripts built around BEDTools (Quinlan and Hall, 2010; Quinlan, 2014), Picard <http://broadinstitute.github.io/picard/> and Samtools (Li et al., 2009) based parsing of the bowtie2 alignment files. Clustering was conducted using a Gaussian Mixture Model (GMM) (Bishop, 2006) and the number of clusters determined through variational Bayesian inference (Corduneanu and Bishop, 2001).

All CONCOCT analyses were completed using a provided pre-configured Docker Image (Merkel, 2014), a form of lightweight distributable process isolation container. This was downloaded from DockerHub <https://hub.docker.com/r/binpro/concoct/> on 2015-10-25.

Additionally, the cut contigs were taxonomically assigned using TAXAssign <https://github.com/umerijaz/TAXAassign> against the NCBI nt database. The BLAST database was downloaded using update_blastdb.pl script

http://www.ncbi.nlm.nih.gov/blast/docs/update_blastdb.pl and TAXAssign was run in parallel (using GNU parallel (Tange, 2011)) with a maximum of 10 reference matches per contig a minimum percentage identity for assignment to a given taxonomic level of 60, 70, 80, 95, 95, and 97 for Phylum, Class, Order, Family, Genus and Species respectively.

CONCOCT clusters were then evaluated using the taxonomic assignments from TAXAssign using the provided “validate.pl” script.

Finally, another attempt at taxonomic assignment using a custom ORF based pipeline was attempted:

- ORFs with a minimum size of 300 were called using Tetrahymena and Universal encodings from contigs over 500bp
- ORFs were then clustered at 90% identity using CD-HIT
- Diamond BLASTP searches were then done against the NR protein database
- Taxonomy was assigned to each contig based on the lowest common ancestor of all its ORFs with hits (via the “lca_mapper.sh” accessory script in MEGAN)
- Contigs were then binned based on the identity of this taxonomic assignment:
 - Endosymbiont contigs were all those assigned to Archaeplastida or descendent node.
 - Host contigs were all those assigned to Aveolata or a descendent node.
 - Eukaryote was a super group containing all Eukaryote assigned contigs
 - Contaminant contigs were all those assigned as Bacterial
 - Viral contigs were all those assigned as Viral sequences.

3.3.2.7 VARIANT CALLING

A variant calling on the binned genomic contigs was attempted to assess the clonality of the endosymbiont. Briefly, the top 10 longest contigs binned as “endosymbiont” were retrieved and Q5 trimmed, error corrected reads aligned to them using bowtie2 and output to BAM files for each library. Library BAMs were then combined using samtools “mpileup” with a minimum mapQ threshold of 5. All potential variants with a mapping depth of 0 were filtered out. SNPs were called from this filtered mpileup file using a custom perl script designed for the the wheat genome project (pers. comms. Hall, Neil). This SNP calling used a coverage cutoff of 10%.

Called SNPs were then visualised and statistics calculated using R.

3.3.3 ENDOSYMBIONT ELIMINATION

CCAP 1660/12 and Yad1g1N1 cultures in NCL media with were treated under the following conditions to attempt to remove the endosymbiont. *P. tetaurelia* were used as a control culture and was given the same treatment.

Paraquat was added at both $1\text{mg}\mu\text{l}^{-1}$ and $0.5\text{mg}\mu\text{l}^{-1}$ concentrations. Cultures were maintained under normal 12:12 lit:dark conditions at 15°C . Cultures were inspected daily using light microscopy and assessed for “bleaching”.

Cycloheximide was added to cultures at both $1\text{mg}\mu\text{l}^{-1}$ and $10\text{mg}\mu\text{l}^{-1}$, again cultures were maintained under standard 12:12 lit:dark condition and 15°C . Cultures when looking clear were subcultured and resuspended in NCL without cycloheximide.

Cultures were maintained in the dark without a lit phase at 15°C and inspected every 2 weeks for clearing. This was to prevent providing too much light and further encouraging endosymbiont growth.

3.4 RESULTS

3.4.1 ITS2 PHYLOGENY

The ITS2 phylogeny demonstrates a clear and well supported relationship in all samples between the CCAP 1660/12 and CCAP 1660/13 endosymbionts and the species described as *M. reisseri* (see fig. 3.4.1). Additionally, the Yad1g1N endosymbiont is positively identified as *C. variabilis*. This phylogeny in general is consistent with previous ITS2 phylogenies (Hoshina et al., 2010).

The ITS2 sequences of the CCAP 1660/12 culture demonstrated a variety of SNPs but never more than a single SNP difference from the basal *M. reisseri* polytomy. These SNPs were grouped into 3 categories: 4 different SNPs that were not found in the reverse complement and therefore represent likely sequencing error (1660-13-purified-K4, K8, K6 and K7), 3 different SNPs that were found in both forward and reverse sequencing (1660-12-B6, 1660-12-19 and 1660-12-18) and therefore represent either true diversity or PCR error and finally 1 SNP that was found in forward and reverse sequencing and in two separate PCR reactions from different biological replicates (1660-12-A7 and 1660-12-6). This featured a single base change from A to G (see fig. 3.4.2) at position 126 in the full masked alignment. This SNP could be the result of intranuclear variation of the ITS2 in the multicopy rDNA array.

With the exception of these SNPs these sequences were identical to 3 from previously sequenced *M. reisseri* endosymbionts, specifically CCAP 211/83 culture with *P. bursaria* Pbi host (AB206547.1), the SW1-ZK symbiont from a *P. bursaria* PB-SW1 host (AB506070.1), and TP-2008b from the SAG241.80 culture (FM205851.1) (see fig. 3.4.1).

This polytomy as the sister clade to other *Micractinium pusillum* taxa was highly supported in both ML and Bayesian phylogenies (91.3% of bootstraps and with a posterior probability of 1.00). There was similarly high support for the separate branching of these sequences from the clade containing the *C. variabilis* and *C. vulgaris* endosymbionts (87.3%/1.00) and the existence of a clade comprising these 3 endosymbionts to the exclusion of any *Coccomyxa* sequences was well supported (89.3%/0.95).

Sequences from the Yad1g1N culture formed a clade with high support with other *Chlorella variabilis* species including NC64A. This support the identification of the endosymbiont in this culture as *C. variabilis* 1N.

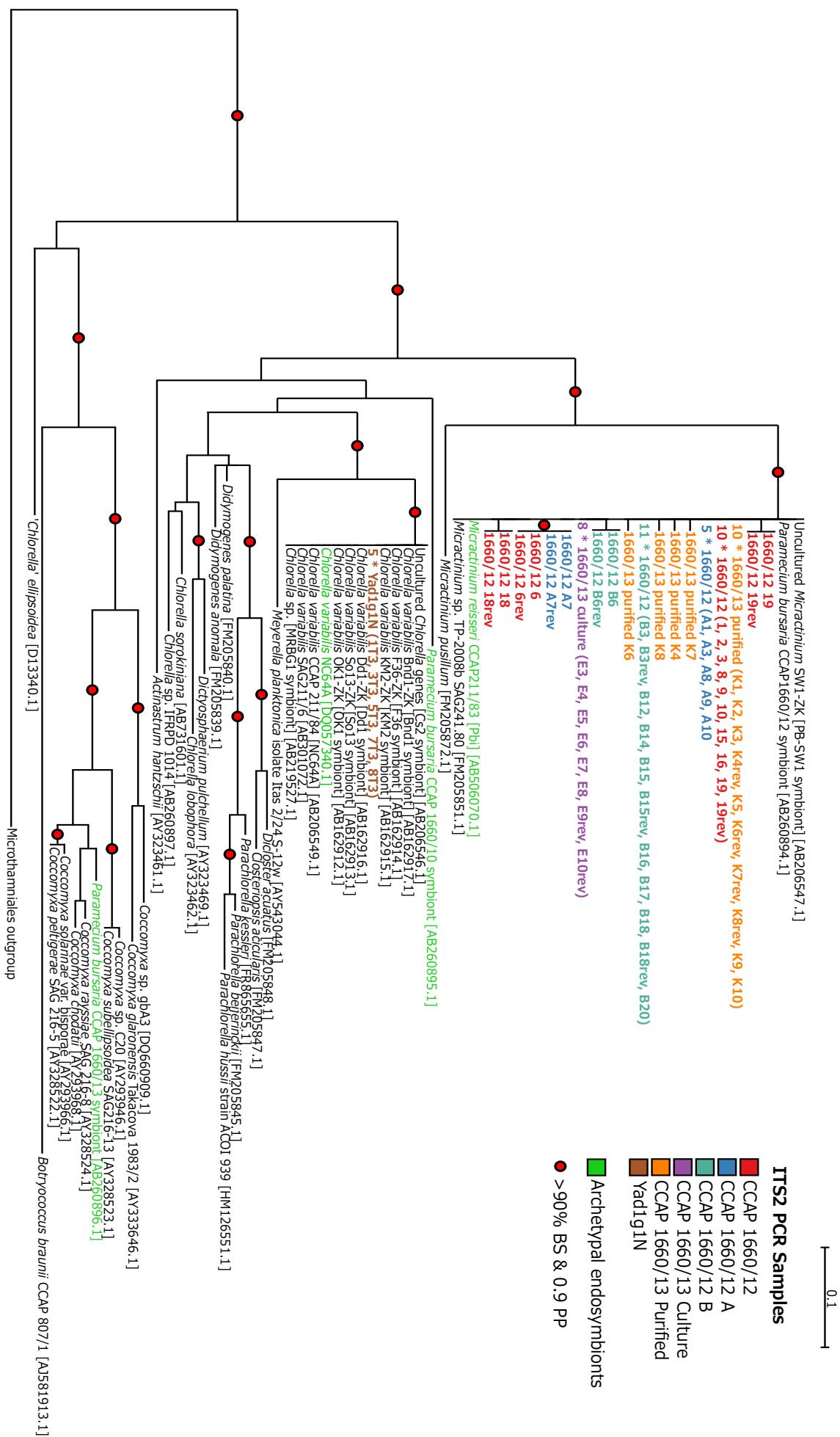


Figure 3.4.1: Combined MrBayes and RAxML phylogeny of all ITS2 sequences along with numerous reference ITS2 sequences from (Hoshina et al., 2010; Hoshina and Fujiwara, 2013). This phylogeny highlights a single example of each of the 4 major groups of *P. bursaria* green algal endosymbionts i.e. *Coccomyxa*, *Microactinium reisseri*, *Chlorella vulgaris* and *Chlorella variabilis*. Multiple clones forming a polytomy have been collapsed into a single node. The highlighted isolates (i.e. 1660/12, 1660/12 A, 1660/12 B, 1660/12 C, 1660/13, 1660/13 Culture, 1660/13 Purified, and Yad1g1N) are highlighted in red.

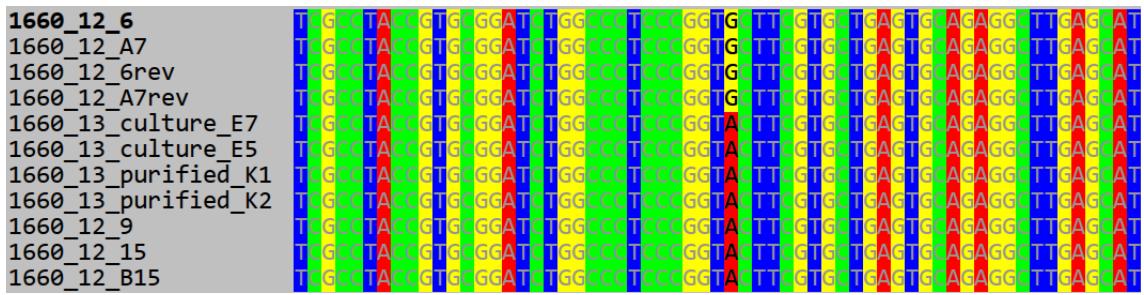


Figure 3.4.2: Alignment showing the sole SNP (at pos 126 in masked ITS2 alignment) that is likely to represent true diversity. This indicates that the endosymbiont population is largely clonal with a small marginally divergent sub-population that has possibly arisen during the endosymbiosis itself. Alternatively, this represents intranuclear variation of the ITS2 across the genomic copies.

| Sample | Raw PE Reads | Q ₃₀ Trimmed PE Reads | Q ₅ Trimmed PE Reads |
|--------|--------------------|----------------------------------|---------------------------------|
| Pb-3 | $3.523 \cdot 10^7$ | $1.951 \cdot 10^7$ | $2.737 \cdot 10^7$ |
| Pb-4 | $3.228 \cdot 10^7$ | $2.606 \cdot 10^7$ | $3.035 \cdot 10^7$ |
| Pb-6 | $3.291 \cdot 10^7$ | $2.437 \cdot 10^7$ | $2.962 \cdot 10^7$ |
| Pb-7 | $4.023 \cdot 10^7$ | $2.642 \cdot 10^7$ | $3.404 \cdot 10^7$ |
| Pb-8 | $3.869 \cdot 10^7$ | $2.613 \cdot 10^7$ | $3.246 \cdot 10^7$ |

Table 3.4.1: Summary of the number of surviving reads for Q5 and Q30 trims in each library

3.4.2 SINGLE CELL GENOMES

3.4.2.1 SEQUENCING AND PRE-PROCESSING

The number of remaining reads in each library after trimming at a minimum average sliding window quality threshold of 30 and 5 can be found in table 3.4.1.

After error correction the combined Q₃₀ trimmed libraries comprised $1.218 \cdot 10^8$ paired end reads. Similarly, the Q₅ trimmed libraries comprised $1.538 \cdot 10^8$ reads.

3.4.2.2 ASSEMBLY

Assemblies were compared using generated contigs and QUAST. Assembly statistics were tabulated to allow comparison (table 3.4.2). As we are interested in recapitulating as much genomic sequence as possible from this complex metagenome but not necessarily to generate “clean” polished closed genome assemblies, the fact that the Q₃₀-SPAdes assembly generated both the longest total assembly (over twice the size of the nearest assembly even when considering only contigs over 1kbp) as well as the highest N₅₀ and within 2kbp of the longest contig of all assemblies (generated by Q₅-SPAdes) was compelling.

Generally, the SPAdes assemblers out-performed Platanus and MegaHit, likely due to being specifically designed for MDA based data. Note, that all assemblies were completed with BayesHammer corrected reads so the difference in performance cannot be attributed to this aspect of the assembly pipeline.

Plots of assembly GC (fig. 3.4.3), cumulative length (fig. 3.4.4) further support Q₃₀-SPAdes as both the longest assembly but an assembly with similar GC profile to the other assemblies and contig length distribution.

| Assembly | textbf{Q30-MegaHit} | Q30-Platanus | Q30-SPAdes-Careful | Q30-SPAdes | Q5-SPAdes |
|--------------------------------|---------------------|---------------------|---------------------------|--------------------|------------------|
| # contigs (≥ 0 bp) | 131,057 | 25,789 | 73,698 | 127,976 | 94,384 |
| # contigs (≥ 1000 bp) | 14,960 | 486 | 13,301 | 21,808 | 12,614 |
| Total length (≥ 0 bp) | 73,350,696 | 6,289,036 | 73,691,706 | 142,281,712 | 81,478,234 |
| Total length (≥ 1000 bp) | 28,064,499 | 2,191,750 | 52,704,642 | 105,269,748 | 58,162,565 |
| # contigs | 41,221 | 776 | 24,923 | 42,180 | 24,109 |
| Largest contig | 13847 | 78624 | 207156 | 207157 | 209,873 |
| Total length | 46,095,605 | 2,398,106 | 60,731,204 | 119,241,116 | 66,064,486 |
| GC (%) | 38.81 | 33.68 | 37.78 | 37.85 | 39.27 |
| N ₅₀ | 1,246 | 6,386 | 4,949 | 7,163 | 6,334 |
| N ₇₅ | 769 | 2,875 | 1,845 | 2,277 | 2,188 |
| L ₅₀ | 10,444 | 112 | 2,937 | 3,530 | 2,241 |
| L ₇₅ | 22,405 | 249 | 7,974 | 11,103 | 6,716 |

Table 3.4.2: Assembly statistics generated by an analysis of contigs using QUAST. Best values are highlighted in bold. All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs (≥ 0 bp)" and "Total length (≥ 0 bp)" include all contigs). N₅₀ and N₇₅ are the minimum contig length at which all contigs of that length are larger comprise 50% and 75% of the total assembly size. Similarly, L₅₀ and L₇₅ are the number of contigs that are summed for a given N₅₀ and N₇₅ (i.e. lower is better). The highest values for each metric across the assemblies is emphasised in bold. This table shows that Q30 Platanus assembly generated the fewest and longest contigs overall, however the Q30-SPAdes assembly generated the longest assembly by a considerable margin with the highest N₅₀.

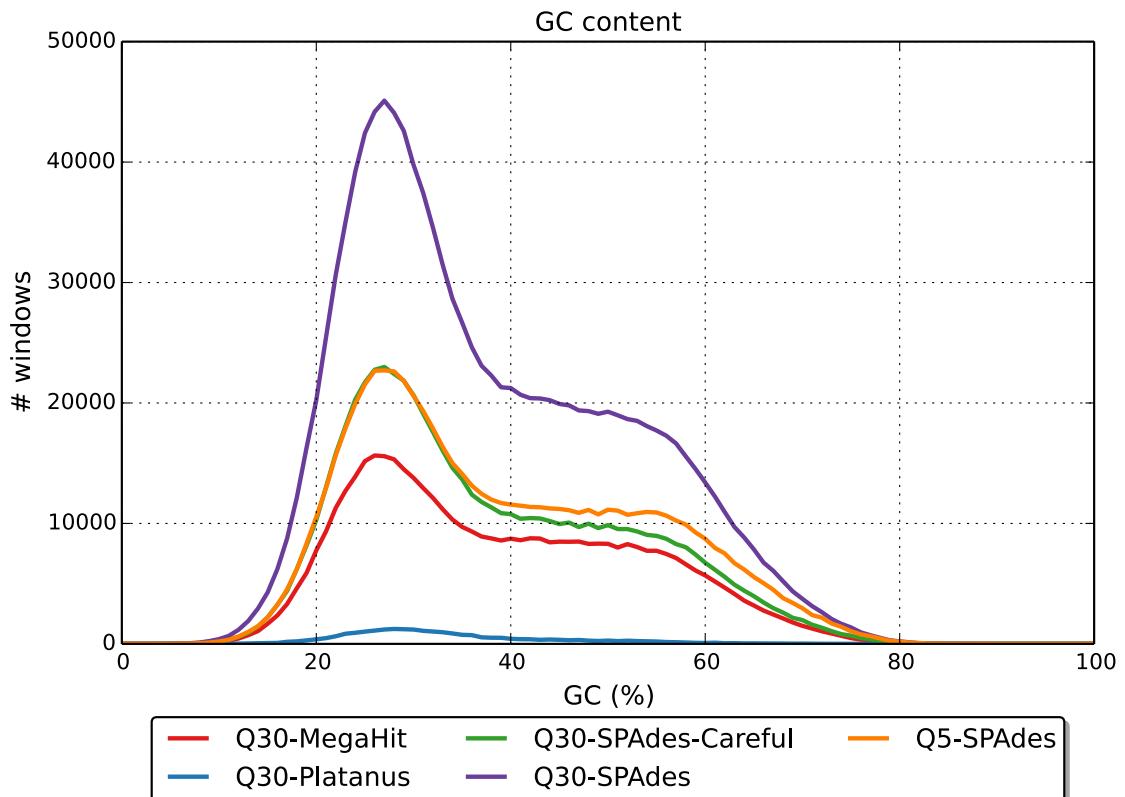


Figure 3.4.3: GC densities of the compared genome assemblies. As expected all display a clear peak around 30% representing that the majority of the assemblies by length contigs are likely to be derived from the GC rich. The height of the Q30-SPAdes peak reflects the relative size of this assembly. Peaks around 50% GC may reflect endosymbiont contigs and possibly bacterial contamination.

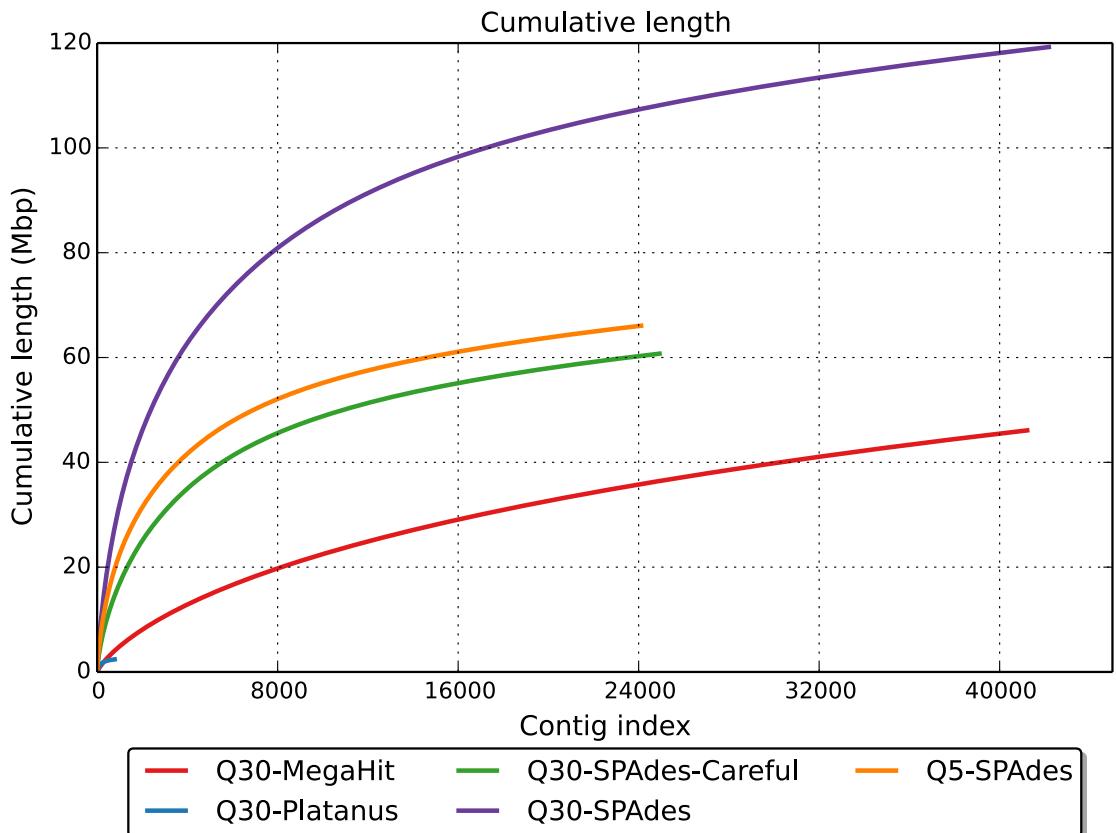


Figure 3.4.4: The cumulative length of contigs as a function of contig number. Again, this plot reflects that Q30-SPAdes generated the largest assembly by a considerable margin and while generating clean consistent contigs Platanus failed to recover many contigs found in other assemblies.

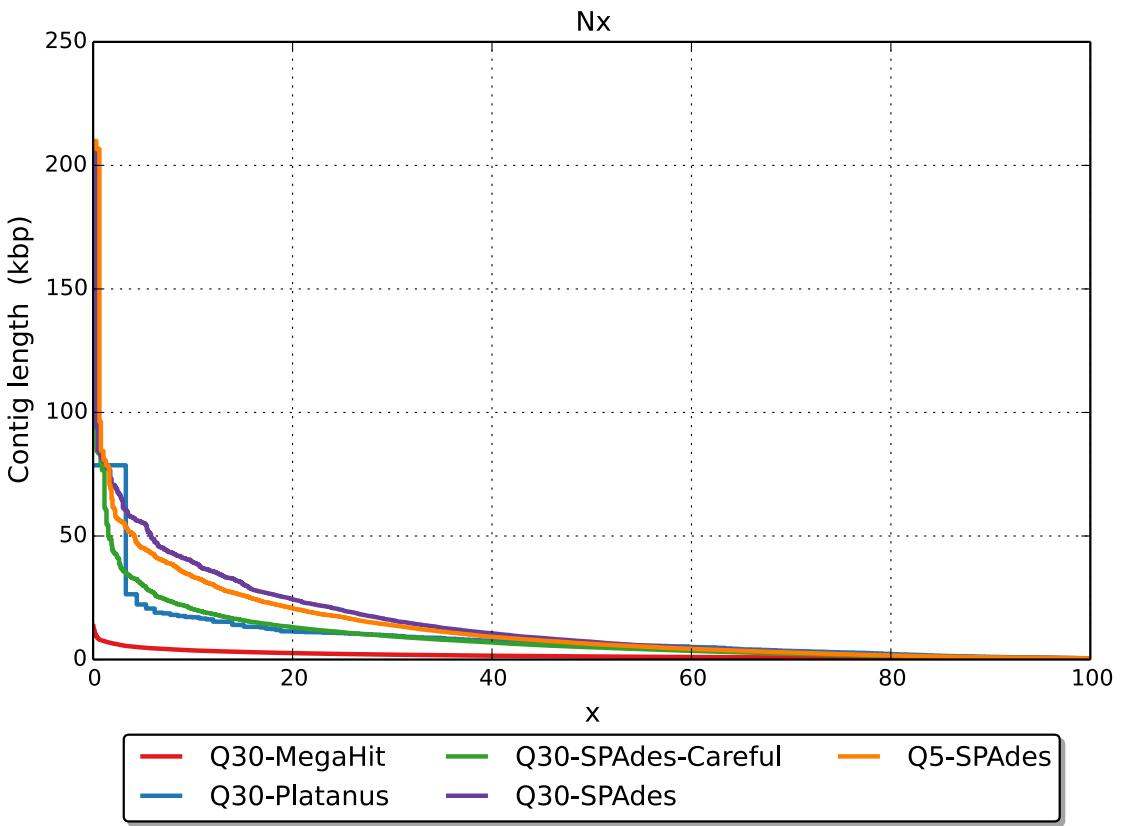


Figure 3.4.5: The number of X (i.e. gap) in the assembled contigs as a function of their length. This demonstrated that generally few Xs were assembled - however, it should be noted that these are contigs and not assembly scaffolds and thus fewer Xs would be expected.

Finally, the plot of Xs indicates that Q₃₀-SPAdes isn't merely a highly gapped assembly (fig. 3.4.5).

Therefore, Q₃₀-SPAdes assembly was selected for further analysis and size filtered to exclude all contigs shorter than 500bp to give 21,090 contigs.

3.4.2.3 BINNING

From the selected Q₃₀-SPAdes assembly, the 21,090 contigs were cut to 10kb fragments for decomposition to generate 64,852 contigs. 18,277 of these 64,852 contigs were successfully given a phylum level assignment, table 3.4.3.

Contigs were clustered into 34 unique clusters by Concoct. These taxonomic assignments were then used to validate the 34 contig clusters generated in Concoct (visualised in fig. 3.4.6) by considering them as the “ground-truth”.

Recall that Precision and Recall can be defined as follows: $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$ where TP are True Positives and FP and FN are False Positives and Negatives respectively (see table 3.4.4 for an explanation of what these terms mean in the context of clustering).

CONCOCT assigned clusters were relatively precise 0.912608 therefore there were relatively few FP i.e. the majority of clusters contained contigs with the same taxonomic assignments.

| Source Group | Number of Contigs | Total Length | Phylum-Level Breakdown |
|---------------------------------|-------------------|--------------|------------------------------|
| Host | 13 | 38,209 | Intramacronucleata |
| | 2 | 140 | Apicomplexa |
| | 1 | 163 | Colponemidia |
| Endosymbiont | 12 | 2,758 | Chlorophyta |
| | 12 | 3,987 | Streptophyta |
| | 1 | 1,674 | Cyanobacteria |
| Bacterial Contamination | 16,230 | 13,435,718 | Proteobacteria |
| | 468 | 669,751 | Firmicutes |
| | 329 | 135,928 | Actinobacteria |
| | 128 | 68,337 | Bacteroidetes/Chlorobi group |
| | 1 | 241 | Deinococcus-Thermus |
| Eukaryotic Contamination | 605 | 640,915 | Ascomycota |
| | 380 | 206,354 | Chordata |
| | 74 | 38,529 | Arthropoda |
| | 12 | 3,623 | Basidiomycota |
| | 7 | 2,150 | Nematoda |
| | 1 | 61 | Platyhelminthes |
| | 1 | 102 | Cnidaria |
| Unknown | 540 | 345,834 | Unclassified |

Table 3.4.3: Summary of taxonomic assignments via TAXAassign grouped into putative “source groups” reflecting the most probable source of 10kb chunked contigs of that specific taxonomic provenance. Of note, is the disproportionate number of contigs from contaminating sources. Specifically, bacteria such as Firmicutes and potential user contaminant in the form of Chordate assigned contigs.

| - | Positive | Negative |
|-------|---|---|
| True | Contigs with same taxonomic assignment are assigned to the same cluster | Contigs with different taxonomic assignments are assigned to different clusters |
| False | Contigs with different taxonomic assignments are assigned to the same cluster | Contigs with the same taxonomic assignment are assigned to different clusters |

Table 3.4.4: A contextual explanation of True and False Positive and Negatives in the context of contig binning/clustering. Top left indicates what a True Positive (TP) means in this context, bottom left a False Positive (FP). Similarly Top Right explains a True Negative (TN) and Bottom Right a False Negative (FN)

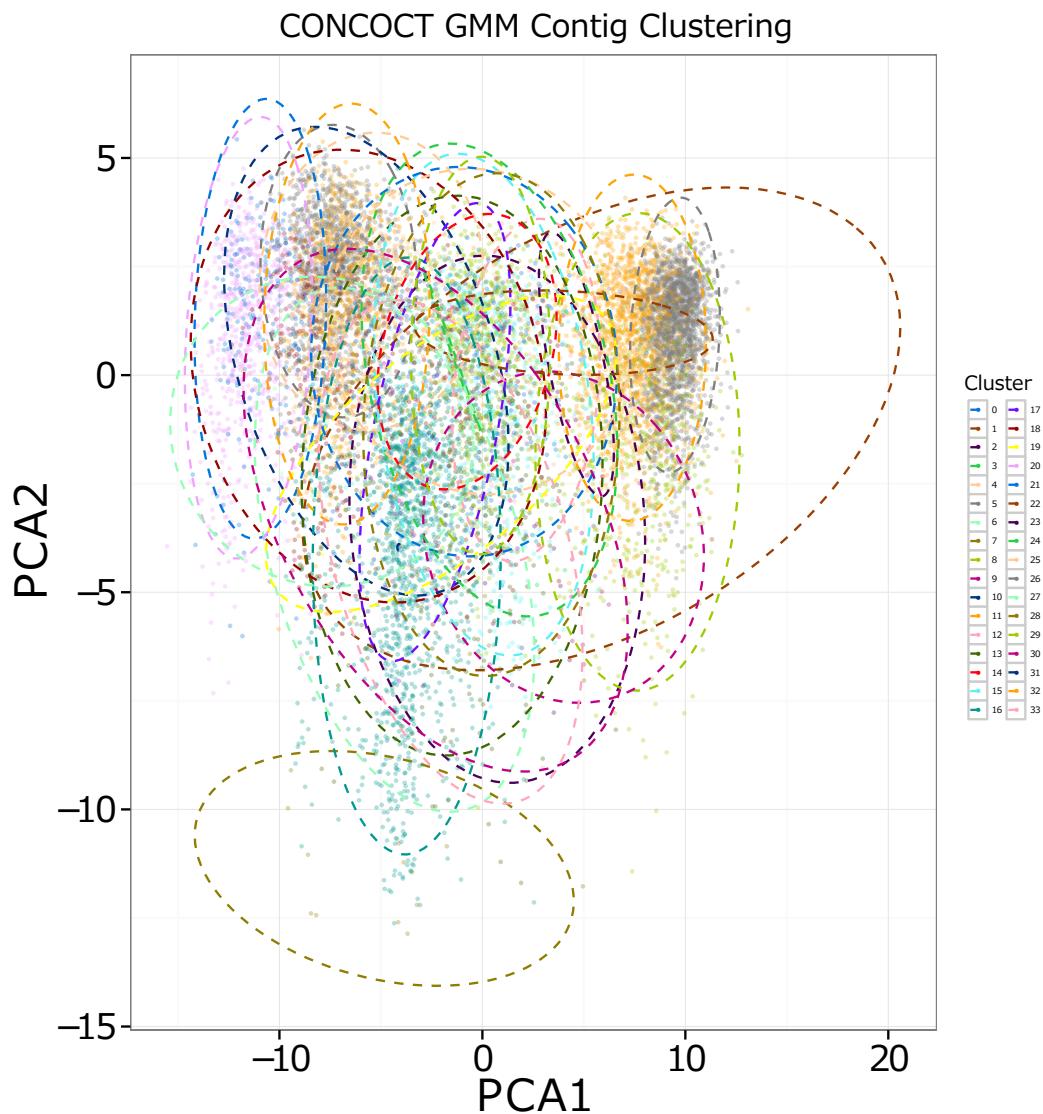


Figure 3.4.6: A low dimensional Principal Component representation of genomic contig cluster assignments. Clusters are assigned via a Gaussian Mixture Model (GMM) based on sequence compositional and coverage features as implemented in CONCOCT. Unfortunately, as can be observed clusters are both poorly distinguished even in the dimensions of the 2 principal components (PCA1 and PCA2) and there are many of them (34). This figure highlights how poorly resolved and noisy the decomposition of this single cell metagenome.

| Bin | Number of Contigs | Total Size (in bp) |
|--|-------------------|-------------------------|
| Endosymbiont Host | 782 3,451 | 1,767,324 24,294,611 |
| Other Eukaryotic Bacterial and Unknown | 1,646 15,211 | 7,237,238 26,342,475 |

Table 3.4.5: Results of the customised taxonomic binning, note the far less conservative binning compared with TAXAssign. Note this only consisted of contigs over 500bp in length

However, recall was relatively poor 0.542250 suggesting a fair number of FN i.e. contigs with the same taxonomic assignments were not confined to a single cluster and were spread over main clusters.

The F_1^1 score for CONCOCT clustering was therefore 0.680389 under the assumption that TAXAssign represents the ground-truth.

It is also worth noting that the 34 clusters had a relatively high level of mutual information (Normalised Mutual Information of 0.332022 and a Rand Index of 0.499741) suggesting many small but highly similar clusters were created. This level of similarity combined with the poor recall suggests a greater number of clusters were inferred than was present in the taxonomic assignment ground-truth. This is likely due to the variational inference of cluster numbers being partially reliant upon sequencing coverage features. As MDA is known to generate very uneven coverage due to amplification biases this likely explains the erroneous clustering.

Therefore, clustering and TAXAssign binning methods were abandoned and the custom ORF based pipeline bins (table 3.4.5) used for variant calling.

3.4.2.4 VARIANT CALLING

The variant calling demonstrated that the majority of potential variants were present in almost all endosymbiont genomes (fig. 3.4.7). Indeed, the highest number of variants (≥ 75) were present in 80 – 90% of endosymbionts.

3.4.3 ELIMINATION OF ENDOSYMBIONT

All elimination analysis focused on the *P. bursaria-M. reisseri* CCAP 1660/12 strain. After 1 week $10\mu\text{g ml}^{-1}$ Paraquat treated CCAP 1660/12 were partially bleached with few visible green *M. reisseri* present in the cells under light microscopy. Unfortunately, after 2 weeks, and despite regular feeding, all CCAP 1660/12 treated with paraquat appeared to die with lysis of the *Paramecium*.

To assess whether this phenotype was due to a too great concentration of paraquat the experiment was repeated at a lower concentration ($1\mu\text{g ml}^{-1}$). Unfortunately, this lead to the same process of gradual bleaching of the CCAP 1660/12 cultures followed by their death. The difference being at the lower concentration this occurred over a 6 week timeframe instead of 2 weeks.

A similar pattern was observed with both concentrations of cycloheximide where $10\mu\text{g ml}^{-1}$ treatment led to a reduction in endosymbiont abundance by 90% after 1 week followed promptly by host death. The lower concen-

¹ $F_1 = 2 * \frac{\text{precision} * \text{recall}}{(\text{precision} + \text{recall})}$

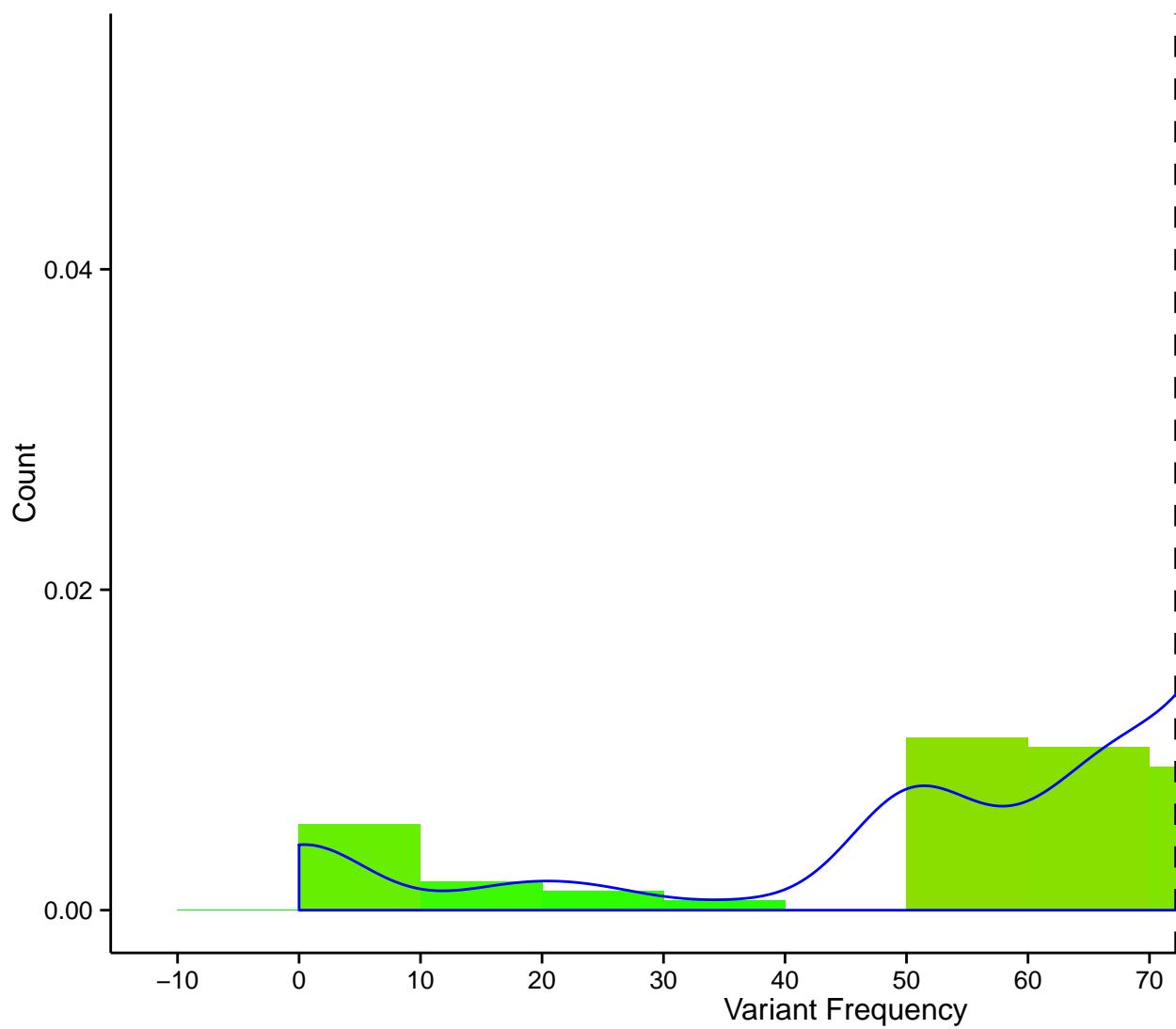


Figure 3.4.7: Histogram summary of variant frequency in the longest 10 custom binned “endosymbiont” contigs (total length 269,755bp). Variant frequency is the percentage of total chromosomes in the sample that share that variant and count axis shows the density (blue line). The vertical dotted line indicates the mean of the sample and the heatmap shows absolute counts in the histogram bins. This figure shows the majority of variants are present in the majority of endosymbiont cells within the host.

tration $1\mu\text{g ml}^{-1}$ displayed the same pattern but over a 6 week period.

Finally, with subculturing and feeding cultures maintained in constant darkness did lead to gradual bleaching over 4-8 weeks. However, after 10 weeks the cultures died with no visible *P. bursaria* cells.

3.5 DISCUSSION

3.5.1 CCAP 1660/12 AND CCAP 1660/13 CONTAIN LARGELY CLONAL *M. reisseri* SYMBIOMTS

Phylogenetic analysis demonstrates that the endosymbiont present in the CCAP 1660/12 and CCAP 1660/13 is *M. reisseri*. The ITS₂ sequences derived from these 2 cultures across 5 different biological replicates (using both primer sets and with or without extra purification steps to minimise contamination from any algae present in the media) were all identical (with the exception of individual SNPs) to 3 separate previously published *M. reisseri* ITS₂ sequences. That this formed a well supported clade with other *Micractinium* sequences and was clearly a distinct grouping from the other *P. bursaria* endosymbiotic green algal species further supports the identity of the 1660/12 and 1660/13 endosymbionts as *M. reisseri*.

While 8 different SNPs were identified in the ITS₂ sequences, these never occurred in the same sequence and half are easily attributable to sequencing error as they couldn't be recapitulated in reverse sequencing of the same clone. Of the remaining 4 SNPs that were validated as not being sequencing error, only 1 was discovered in separate PCR reactions and biological replicates and thus can putatively be attributed to genuine biological diversity and not merely PCR error (ITS₂-6 and ITS₂-A7, A to G transition). Therefore, on the basis of ITS₂ sequences we cannot say the endosymbionts in CCAP 1660/12 and 1660/13 form a clonal population. However, a single SNP in the hypervariable ITS₂ region represent very recent and minor divergence. The most likely explanation is that this represents the emergence of a slightly modified line of endosymbionts within the clonal endosymbiont population of the CCAP 1660/12 culture or intranuclear variation within a single clonal population. The distribution of SNP variants on endosymbiont binned contigs supports this hypothesis. This is because the majority of SNPs were detected to be present in 73% or more of endosymbiont chromosomes. Due to the uniformity of the ITS₂ sequences there is no evidence of multi-strain photobiont co-habitation as described by (Hoshina, 2012).

It should be noted that the majority of ITS₂ based studies make use of the secondary structure (predicted using tools such as RNAstructure (Mathews et al., 2004)) in inference (Schultz and Wolf, 2009). This increases reliability of phylogenetic inference (Keller et al., 2008) allows ITS₂ to be used to distinguish higher taxonomic levels (Coleman, 2003), and plays a role in resolving the thorny problem of species determination (Müller et al., 2007). However, as the endosymbiont species ITS₂ secondary structures have already been extensively investigated (e.g. (Hoshina and Imamura, 2008; Hoshina et al., 2010)) and are generally better suited for analysis of more divergent taxa, it was considered unnecessary to conduct structural analysis for taxonomic analysis of these endosymbionts.

3.5.2 RELIABILITY OF CULTURE COLLECTION

One clear result and point worth raising is that contrary to previous studies (accession AB260896.1 ([Hoshina and Imamura, 2008](#))) and CCAPs culture description CCAP 1660/13 does not contain a *Coccomyxa* endosymbiont and contains an identical *M. reisseri* endosymbiont to the CCAP 1660/12 culture. Unfortunately, on communication with CCAP it emerged that the 1660/12 strains in their collection are no longer available and that CCAP 1660/13 had apparently become overgrown by free-living *Coccomyxa*. Therefore it is likely that the previous finding of *Coccomyxa* “endosymbionts” in CCAP 1660/13 ([Hoshina and Imamura, 2008](#)) represents accidental contamination and sequencing of the free-living *Coccomyxa* also present in the culture.

The identical nature of the CCAP 1660/12 and CCAP 1660/13 endosymbioses is perhaps not surprising when it is emphasised that these cultures were isolated from the same pond (Cambridge, UK) by CCAP.

This demonstrates the necessity of not taking culture collection labels and taxonomic assignments on faith. It is critical to thoroughly determine that all received cultures actually contain the organism

3.5.3 MDA METAGENOMES ARE NON-TRIVIAL

The biases induced by MDA in single cell genomes are known to be formation of chimeric sequences and the amplification of undesired contaminant sequences ([Binga et al., 2008](#)). Additionally, despite a theoretical basis that the amplification coverage bias should be random ([Hosono et al., 2003](#)) there is evidence disputing this in practice ([Ellegaard et al., 2013b](#)). The magnitude of this bias is related to the starting quantity of DNA ([Ellegaard et al., 2013a](#)). Fortunately, there does not appear to be any bias related to GC ([Ellegaard et al., 2013a](#)). An increase in the number of starting cells to the range of a few hundred to a few thousand bacterial cells improves amplification considerably ([Ellegaard et al., 2013a](#)). Unfortunately, increasing the number of cells in the case of CCAP 1660/12 *P. bursaria* - *M. reisseri* system would likely compound issue with bacterial contamination due to both a greater sample volume leading to greater inclusion of food bacteria living in the media and an increase in the number of partially digested bacterial (and viral) symbionts associated with the host.

SPAdes, by far, generates the best assemblies of complex MDA-based metagenomes of the assembly tools trialled. This cannot be attributed to the effective read error correction implemented as part of SPAdes via BayesHammer as all assemblies were completed on BayesHammer error corrected reads. The performance of SPAdes is likely attributable to 2 factors: it is specifically designed to handle MDA-based single cell assemblies and thus is highly tolerant of the coverage variability observed and secondly it is the lone genome assembly that effectively utilised paired-end data during assembly. The vast majority of assemblers will only utilise this data in ad-hoc post-assembly heuristic operations to improve contigs and scaffold the dataset. On the other hand, SPAdes generates the assembly de-Bruijn using siamese rectangular graphs that incorporate both forward and reverse reads and their respective insert. In future, it may be worth re-analysing this data using other MDA-specific tools such as HyDA to assess their performance.

The relative performance of Q₃₀-SPAdes with and without the “careful” setting is interesting. This setting minimises the risk of mismatch and indels found in the assembly. This led to assembly with statistics relatively

similar to that of the Q5-SPAdes assembly. However, on correspondence with the developers of SPAdes it emerged that there was a bug in this setting in the version of the assembler used within this study leading it to be highly conservative and discard many assembled contigs that were unlikely to be mismatches.

Finally, the poor performance of CONCOCT suggests that coverage and composition are not effective metrics by which to decompose an MDA-based metagenome into constituent “bins”. The poor recall and high similarity indices between the clusters suggests that a greater number of clusters were inferred than was present in the ground truth of the taxonomic assignments. This likely represents the effect of biased amplification in MDA (therefore heterogeneous variable coverage) on the variational inference of the number of clusters and the utility of the coverage feature in general. This means, therefore, in MDA-based metagenomes standard metagenomic binning pipelines that are reliant on coverage metrics (even partially as in the case of CONCOCT) are not effective.

This problem is somewhat symptomatic of the current state of the tool ecosystem for MDA-based eukaryotic metagenomes. The few MDA-orientated analysis tools focus on the assembly of bacterial systems whereas the majority of the metagenomic tools are based on features and metrics such as coverage that are only consistent in conventional non-MDA bulk genomic studies. Ideally, future research will improve the ease of analysis and assembly of datasets such as this.

3.5.4 METABOLIC CO-DEPENDENCE IN THE CCAP 1660/12 SYSTEM

Due to the repeated failure to create endosymbiont free *Paramecium* hosts from the CCAP 1660/12 cultures using 3 of the major accepted methodologies (cultivation in darkness (Karakashian, 1963), paraquat (Hosoya et al., 1995; Tanaka et al., 2002) or cycloheximide (Weis, 1984)) we are forced to address the possibility that the *Micractinium reisseri* endosymbiont and *P. bursaria* system in CCAP 1660/12 and CCAP 1660/13 cultures forms an obligate system. By some unidentified mechanism, metabolic co-dependence may have become fixed in this culture.

Cycloheximide does partially inhibit host protein synthesis (Weis, 1984; Kodama et al., 2007; Kodama and Fujishima, 2008, 2009) therefore it is possible that in the host strain found in the CCAP 1660/12 culture that this partial inhibition is lethal to both host and endosymbiont. However, the failure of this method in conjunction to Paraquat, a herbicide which theoretically should only affect the endosymbiont, and constant dark culturing suggests it is the loss of endosymbiont photosynthetic activity that is lethal to the host cells (as adequate bacterial foodstocks were included in these cultures).

The one major method that wasn’t attempted was the use of 3-(3,4-dichlorophenyl)-1,1-dimethylura (DCMU) an established blocker of photosystem II (van Gorkom, 1974). However, DCMU has previously been found to be mildly toxic in *P. bursaria*, affecting the sexual reproduction system (Miwa, 2009) therefore, this would have proven unlikely to show different results in either the case of a particularly “sickly” host strain or obligate endosymbiosis.

This result indicates the presence of key differences between the current state of this endosymbiosis and the previously studied *C. variabilis* endosymbiosis studied by (Kodama and Fujishima, 2014). Therefore, a comparative analysis of these systems could theoretically shed light on the mechanism by which metabolic co-dependence

has become fixed in one system. Alternatively, this difference may just reflect the nature of two different, independently acquired endosymbioses with different species and strains of both host and endosymbiont.

Another avenue of study that we did not investigate was that of isolation of endosymbiont into free-living cultures (Achilles-Day and Day, 2013a). This would allow us to establish whether green algae such as *Micractinium* that have obligate hosts are themselves obligate endosymbionts. There is some evidence pointing towards this in nature, with the widespread predation of *M. reisseri* and *C. variabilis* by their specific PBCV viotypes as well as the relative paucity of natural free-living strains of these species. To my knowledge, there has only been a single isolated and characterised free-living *M. reisseri* (Abou-Shanab et al., 2014) example and no *C. variabilis* examples. However, this said, algae have previously been isolated from the CCAP 1660/13 culture (Achilles-Day and Day, 2013a). We have demonstrated via ITS₂ sequencing that the endosymbionts in CCAP 1660/13 are the same as those in CCAP 1660/12. Therefore, if these isolated algae are actually endosymbionts (as supposed to the free-living *Coccomyxa* sp. that overgrew the culture shortly after this study was published) then the *M. reisseri* endosymbiont is capable of living without the host and is not an obligate endosymbiont despite *P. bursaria* being an obligate host.

3.6 CONCLUSIONS

Therefore, on the basis of ITS₂ sequencing the CCAP 1660/12 culture endosymbiont is a strain of *Micractinium reisseri*. Additionally, the CCAP 1660/13 endosymbiont has been misclassified as a strain *Coccomyxa* and is the same *Micractinium reisseri* species found in the CCAP 1660/12 culture. I have confirmed the Yad1g1N endosymbiont as being *C. variabilis* 1N. Despite poor performance in genome assembly, the evidence of the genomes and ITS₂ data seem to indicate that this endosymbiont forms a clonal or near clonal population within the CCAP 1660/12 endosymbiont. Similarly, on the basis of ITS₂ diversity the Yad1g1N culture contains a clonal 1N endosymbiont population. At a minimum, a single strain of *M. reisseri* comprises the sole green algal endosymbiont in the CCAP 1660/12 and 1660/13 cultures although there may be intranuclear variation of the ITS₂ or it may be actively evolving as evidenced by a small divergent sub-population.

Finally, the *Micractinium reisseri* endosymbiont in cultures CCAP 1660/12 CCAP 1660/13 has potentially become metabolically co-dependent with the host. The host appears incapable of survival without the endosymbiont, therefore, it is important to attempt to identify the differences between the demonstrably facultative relationship between the Japanese Yad1g1N strains (used in (Kodama and Fujishima, 2014)) and the putatively obligate CCAP 1660/12. Identifying these differences may pinpoint the mechanism by which metabolic co-dependence becomes fixed in *P. bursaria* - green algal endosymbioses.

"even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience;"

- David Hume: *A Treatise of Human Nature*, 1738

4

Transcriptomic analysis of the *Paramecium bursaria* and *Micractinium reisseri* endosymbiosis

4.1 INTRODUCTION

The *Paramecium bursaria*-*Micractinium reisseri* (PbMr) endosymbiosis conveys phototrophy (Karakashian, 1963), numerous photobiological traits (e.g. (Berk et al., 1991; Saji and Oosawa, 1974; Nakajima and Nakaoka, 1989; Niess et al., 1982b; Iwatsuki and Naitoh, 1988; Summerer et al., 2009), partially reviewed in (Sommaruga and Sonntag, 2009)) and its establishment and maintenance is dependent on photosynthetic activity and enigmatic light-induced factors (Karakashian, 1963; Hosoya et al., 1995; Kodama et al., 2007; Kodama and Fujishima, 2014). Therefore, a relatively unbiased global metatranscriptomic profile of host and endosymbiont in both lit and dark conditions would potentially identify key transcripts which play a role in the establishment, maintenance, and characteristics of this endosymbiosis.

"Dual-RNAseq" is a form transcriptomics which characterises transcripts in a small number of defined organisms simultaneously (Westermann et al., 2012). It has proven an effective method in several studies investigating host-chloroplast interactions (Nowack et al., 2011; Jiggins et al., 2013; Xiang et al., 2015), and host-pathogen systems (Tierney et al., 2012; Kawahara et al., 2012; Jones et al., 2014; Hayden et al., 2014). It differentiates itself from both standard metatranscriptomics, such as those common in microbial ecology (Poretsky et al., 2005;

(Aliaga Goltsman et al., 2014), by being conducted on samples of known, or mostly known composition, and from classical transcriptomics by not depending on axenic samples.

Paramecium bursaria and its green algal endosymbionts form a system well-suited for “dual-RNAseq” analysis. Firstly, there is a plethora of literature on the physiology and behaviour of host and endosymbiont, both together and individually (e.g. (Iwatsuki and Naitoh, 1988), see (Kato and Imamura, 2009b) and the Introductory Chapter for more details), presenting a key resource by which results can be contextualised. Additionally, transcriptomic analysis has proven feasible in reasonably close relatives of both host (Arnaiz et al., 2010; Kolisko et al., 2014) and endosymbiont (Guarnieri et al., 2011; Rowe et al., 2014; Bashan et al., 2015). Even more promisingly, there has been an analysis of the host-endosymbiont system (although in a different strain: Yadi1N) (Kodama et al., 2014). Unfortunately, this study focused only on the expression pattern of the host alone with and without its endosymbiont and discarded endosymbiont derived data during analysis.

This said, the PbMr system does also present some severe difficulties in terms of its transcriptomic tractability. Specifically, the system is highly genetically and transcriptomically complex with *P. bursaria*’s ciliate nuclear dimorphism and high order polyploidy (Raikov, 1995), the presence of sexual reproduction in both host (Jennings, 1939) and endosymbiont species (Blanc et al., 2010), and a large range of GC biases (Kodama et al., 2014). Therefore, care must be taken to optimise sequencing, and assembly methods to mitigate these complications.

These difficulties are compounded by the lack of available reference genomes for either *Paramecium bursaria* or *Micractinium reisseri* and thus necessitating *de novo* transcriptome assembly. However, the utility of sequenced genomes from divergent ciliate species (i.e. *Tetrahymena thermophila* (Eisen et al., 2006), *Paramecium tetraurelia* (Aury et al., 2006) and *Paramecium caudatum* (McGrath et al., 2014)) and endosymbiotic green algae *Chlorella variabilis* NC64A (Blanc et al., 2010) and *Coccomyxa subellipsoidea* C-169 (Blanc et al., 2012) (see fig. 1.2.2 in the Introductory Chapter and fig. 3.4.1 in Chapter 3 for respective phylogenetic context of these genomes) as references for assembly was investigated. It should also be noted that the existing *Paramecium bursaria* (Kodama et al., 2014), *Paramecium duboscqui* (Kolisko et al., 2014) and *Chlorella vulgaris* (Guarnieri et al., 2011) transcriptomes mentioned above were successfully recapitulated *de novo* (without a reference genome).

The mixotrophic nature of the host *Paramecium* (Dolan, 1992) means there are partially digested bacterial prey species, as well as numerous associated bacteria (Görtz and Fokin, 2009; Fokin and Görtz, 2009; Schrallhammer and Schweikert, 2009) and viruses (Van Etten et al., 1983) which all present potentially obfuscating sources of contamination in the analysis of host-endosymbiont interaction. Therefore, it is key to effective analysis of this system to develop methods that minimise the effects of contamination at all stages of analysis. To address this, we investigated methods to reduce contamination during library preparation such as washing steps, cell picking and single cell sequencing techniques; methods to screen and/or filter sequenced libraries for contaminants before inclusion in assembly and methods to effectively sort assembled transcripts into bins relating to their likely originating organisms (i.e. “host”, “food” or “endosymbiont” derived).

To this end, bulk RNAseq libraries from cultured PbMr was sequenced using 76bp paired-end reads and the Illumina Gene Analyzer II platform taking care to minimise contamination by filtering and washing cultures and carefully assessing culture health to maximise the number of healthy PbMr sequenced. Unfortunately, due to

limitations in the maintainable culture density of the *Paramecium bursaria* CCAP 1660/12 and thus the quantity of extractable mRNA it was necessary to pool all day and night replicates into a single pair of day and night libraries.

While this provided sufficient material for sequencing it precluded accurate inference of differential expression between day or night by masking the biological replicates (Auer and Doerge, 2010). We, therefore, also sequenced a set of 3 (followed later by an additional 5) dark and 3 light biological replicates using single-cell RNAseq (sc-RNAseq) methods. This also allowed a finer-grain control over cell selection and potentially a method to reduce culture based contamination.

While reasonably nascent, sc-RNAseq has shown a lot of promise in well characterised systems such as human cell cultures (Bengtsson et al., 2005; Shalek et al., 2013) and *Saccharomyces cerevisiae* (Lipson et al., 2009) and there are high expectations of their utility for “dual-RNASeq” (Westermann et al., 2012). sc-RNAseq addresses the key difficulties of analysing unculturable or poorly culturable organisms (Murray et al., 2012) and investigating cell-cell heterogeneity in expression patterns (Raj and van Oudenaarden, 2008; Shalek et al., 2013)). Uninvestigated, this heterogeneity (either from biological and/or genomic variance or just the stochasticity of gene expression), can lead to a Yule-Simpson effect (Yule, 1903; Simpson, 1951), where the false amalgamation of distinct expression patterns in previously cryptic but distinct cellular subpopulations could generate a spurious expression pattern contrary to either subpopulation.

There are a range of possible sc-RNAseq methods (whose advantages and disadvantages are covered in the Methods Chapter). We used Qiagen’s Repli-G Whole Transcriptome Amplification (WTA) MDA-based kit as MDA is well established and characterised in single cell genomics (e.g. (Spits et al., 2006)), has a simple methodology not requiring additional equipment, and is potentially more successful at recovering transcripts from a wide range of abundance levels than other methods i.e. recovers many lowly expressed transcripts¹. Unfortunately, despite the publication of empirical comparisons of single cell transcriptomic methods (Wu et al., 2014), Qiagen’s Repli-G WTA MDA-based kit has yet to be directly assessed relative to other approaches and thus its performance has not been independently verified. Briefly, this method involves the ligation of reverse transcribed cDNAs using oligo-dT primers (after lysis and removal of gDNA) before MDA by a φ29 DNA polymerase with a 5'-3' exonuclease proofreading activity (reducing error-rate of amplification to $9.5 \cdot 10^{-6}$ errors per nucleotide (Paez et al., 2004) compared with 10^{-4} to 10^{-5} for Taq (Tindall and Kunkel, 1988; Eckert and Kunkel, 1990)) (Korfhage et al., 2015).

Unfortunately, despite its utility sc-RNAseq generates a new set of difficulties. First and foremost, there has only been a single published use of sc-RNAseq, to my knowledge, in non-model unicellular eukaryotes. This study by (Kolisko et al., 2014), briefly addressed the issues of bias, contamination and gene discovery effectiveness in a set of model and non-model eukaryotes and constitutes an important proof-of-concept. However, it also used a different sc-RNAseq approach (SMART), focused on single organisms, and didn’t address, in-depth, the optimal way to process, assemble and utilise single cell datasets from protists. While some work has been done investigating the optimal pre-processing of bulk RNAseq datasets (e.g. (Macmanes and Eisen, 2013; Macmanes,

¹ <https://www.qiagen.com/gb/shop/sample-technologies/rna-sample-technologies/total-rna/repli-g-wta-single-cell-kit/> as of 2015/08/25

2015) the effect of different trims and error correction on sc-RNAseq has yet to be characterised. There are also some early indications of problems of cryptic bacterial contamination from samples and/or reagents in sc-RNAseq is particularly problematic (Kolisko et al., 2014). This further increases the importance of library screening and post-assembly transcript binning.

4.2 AIMS

Therefore, this chapter will investigate the optimal use of 2nd generation bulk and sc-RNAseq libraries in a characterising a complex reference-free system. Specifically, it will look at the screening of RNAseq libraries for contamination before assembly, the optimal preprocessing (partitioning, trimming, digital normalisation and error correction), assembler and assembly parameters (including the utility of divergent reference genomes from related species) in recapitulation of host and endosymbiont transcripts. Finally, I will address the problem of the attribution of recovered transcripts into their appropriate likely originating organism.

4.3 METHODS

4.3.1 SAMPLE PREPARATION AND SEQUENCING

4.3.1.1 BULK TRANSCRIPTOME RNA PREPARATION

For bulk transcriptomic analyses CCAP 1660/12 cells were harvested in a way to minimise contamination from bacterial prey species in the culture. $\sim 10^6$ cell aliquots were strained through $40\mu m$ sieves, filtered on $10\mu m$ nylon filters, before finally being filtered on $8\mu m$ TETP polycarbonate filters using a low-pressure filtration pump. Collected samples were either immediately quick-frozen in liquid nitrogen for storage ($-20^\circ C$ for short-term storage and $-80^\circ C$ for longer storage) or harvested by centrifugation. In order to investigate the two main metabolic states of the symbiosis (i.e. under light conditions during active photosynthesis and in the dark when no photosynthesis is taking place) samples were extracted 5 hours into the light and dark phase of the 12:12 hour day-night cycle.

To ensure extracted RNA was representative of healthy and interacting host and endosymbionts care was taken to the number of dead/dying cells from which RNA was extracted. In order to do this, a subsample was taken from each culture during the process of harvesting and scored for dead/dying cells. Cell assays were formed by taking 1-2ml of each harvest cell pellet and fixed using $40\mu l$ Lugol's solution (0.5g I_2 and 1g KCl in 8.5ml of MilliQ water). Dead/dying cells were identified as broken or puckered cells and counted using light microscopy. Samples containing $>10\%$ dead/dying cells were discarded and no RNA extracted from them.

In order to lyse collected samples, cells were washed from the filter or the pellet was resuspended in 1ml TriReagent (Sigma) heated to $60^\circ C$. Cells were vortexed with sterile $300\mu m$ glass-beads for 15s, incubated at room temperature for 10 minute, vortexed for 15s, quick-frozen in liquid nitrogen and stored at $-20^\circ C$ before further processing. Samples were defrosted, vortexed for 15s, placed in a heat-block set to $60^\circ C$ for 10 minutes while continuing to be vortexed, removed from heating and vortexed again for 15s. RNA was extracted by adding 0.2ml

of Chloroform to the glass-bead-trizol-sample solution, shaking for 15s, incubating for 5 minutes at room temperature and centrifuging at 12,000g for 15 minutes at 4°C. The upper-phase was then transferred to an RNase-free 1.5ml tube and an equal volume (~ 0.5ml) of isopropanol was added before shaking for 15s. The isolated RNA was then incubated at -20°C for 10 minutes (up to several hours) before being collected as a pellet using a centrifuge at 10,000g for 10 minutes at 4°C (supernatant was discarded). The RNA pellet was then washed with 1ml of 75% ethanol and centrifuged twice at 10,000g for 10 minutes at 4°C with the supernatant being discarded after each centrifugation. The pellet was then dried before being resuspended in 100µl of RNase-free water. The RNA was cleaned further using the Qiagen RNeasy clean-up kit before being assessed for quality using ND-1000 (NanoDrop) and BioAnalyzer (Agilent).

4.3.1.2 SINGLE CELL RNA PREPARATION

For single cell transcriptomics, a “cell-picking” approach was used in which *P. bursaria* cells (from the CCAP1660/12 culture) were inspected on an inverted light microscope before being picked using an orally aspirated drawn-glass Pasteur pipette (Garcia-Cuetos et al., 2012). In order to minimise contamination from food bacteria present in the media these picked cells were washed 3 times by serial transfer to 10µl droplets of sterile NCL media. The washed cell was then transferred to a 10µl droplet of sterile water. Cells were picked 5 hours into both the lit and dark phase of the 12:12 hour day-night cycle identically to the bulk analyses. As cells were picked individually, health status could be exhaustively assessed during picking and therefore the subsampling and scoring method used to check the status of cells in bulk preparations was unnecessary.

cDNA was generated and amplified using the MDA-based Qiagen REPLI-g WTA Single Cell Kit (Korfhage et al., 2015) with additional cell disruption steps. Specifically, cells were transferred from their respective 10µl droplets of sterile water to a PCR tube containing 6µl water and 4µl lysis buffer. Due to the robust chitin cell walls of *M. reisseri* (Kapaun and Reisser, 1995) it was important to ensure thorough cell lysis. Therefore, samples underwent mechanical disruption by bead beating (Sigma, 425 – 600µm, acid-washed) followed by freeze-thaw via submersion in liquid nitrogen for 5 seconds. In order to compare disruption methods, extractions and amplifications were also conducted using just lysis buffer, bead beating and vortexing (i.e. without freeze-thaw), and just the lysis buffer. Samples were then quantified using a ND-1000 (NanoDrop) and as extraction methods produced near identical DNA concentrations the maximal disruptive method of freeze-thaw, beat beating, vortexing and lysis buffer described above was used for further purification and library preparation. The samples were then vortexed for 1 minute before a gDNA removal step.

mRNA was selectively amplified and reverse transcribed to cDNA using poly-A selection (i.e. oligo-dT) primers to prevent amplifying ribosomal sequences. Prior to MDA by a φ29 DNA polymerase cDNA were ligated into long fragments due to lower MDA efficiency for short fragments (Korfhage et al., 2015). This reduces size-dependent amplification bias but could potentially lead to the creation of chimeric transcripts in which paired-reads cross boundaries of adjacently ligated cDNA transcripts. Analysis of this is discussed below.

The amplified cDNA was then purified using a QIAamp DNA mini kit and eluted in 100µl elution buffer. This kit operates by binding the DNA to a QIAamp membrane in a spin column followed by successive washing steps

to remove impurities such as remaining proteins and cations. To create 3 dark cDNA samples (Dark1-2, Dark1-3, Dark1-5) and 3 light samples (Light1-9, Light1-10, Light1-11).

Due to low quantities of eukaryote identifiable reads in the initial 3 sequenced single cell dark libraries a set of additional single cell extractions were conducted. These followed the same protocol as above but also featured an additional final PCR-based screening of synthesised cDNA using primers specific for *Paramecium Bug22* sequence. *Bug22* is a highly conserved ciliary protein found in a large number of organisms including the ciliates (Smith et al., 2005b; Laligne et al., 2010), green algae (Keller et al., 2005; Laligne et al., 2010; Meng et al., 2014), higher plants (Hodges et al., 2011), and animals (Mendes Maia et al., 2014) we used as a marker for *Paramecium* derived cDNA. Primers used were Bug22BFWD "GCATTCTAGACCAATCTGGCTTCTGTCAA" and Bug22BREV "GCATTCGAATTGAGGCTCTAAATCTCTCTCA", under standard PCR conditions. 5 (Dark2-2, Dark2-3, Dark2-6, Dark2-7, Dark2-8) samples with bands of appropriate size were then taken forward for library preparation and sequencing.

4.3.1.3 LIBRARY PREPARATION

For both bulk and single cell preparations each cDNA sample was fragmented in $130\mu l$ 1xTE buffer on the Covaris E220 with a target size of 225bp (duty factor of 10%, 200 cycles per burst, peak incident power of 175, 200s at 7°C). Fragment sizes were checked on a BioAnalyzer (Agilent) 7500 DNA chip. cDNA was then concentrated using a GeneRead kit column with a elution in $35\mu l$. Fragmentation step was then repeated 3 times (110s) until majority of cDNA in each library was between 200 – 250bp

cDNA ends were then end-repaired, adenylated and adapters ligated using the NEXTFlex (Biooscientific) sequencing kit according to the manufacturer's instructions and using NEBNext (New England Biolabs) indices. Also following the NEXTFlex kit instructions, MgNa bead purification was done before and after PCR amplification using NEBNext reagents. Finally, prepared libraries were size selected using a Blue Pippin machine at a size selection of 350bp (range 315 – 385bp).

A final bioanalyzer step was conducted with individual library concentrations ranging from $0.66 - 4.09\text{nM}$.

4.3.1.4 SEQUENCING

The bulk day and night library were paired-end (PE) 76bp sequenced using an Illumina Genome Analyzer II by the Exeter University Sequencing Service. The two libraries were sequenced on separate flowcells (Bulk-Light, Bulk-Dark).

Single cell libraries were paired-end 150bp using an Illumina HiSeq 2500 by Exeter Sequencing Service. 3 dark (Dark1-2, Dark1-3, Dark1-5) and 3 light (Light1-9, Light1-10, Light1-11) samples were multiplexed sequenced on a single flowcell lane. The 5 additional dark samples (Dark2-2, Dark2-3, Dark2-6, Dark2-7, Dark2-8) were multiplexed and sequenced on a single flowcell lane in a separate sequencing run.

4.3.2 LIBRARY CONTAMINATION SCREENING

4.3.2.1 TAXONOMIC ANALYSIS

Sequenced libraries were initially screened using the standard metrics implemented in the FastQC to check for standard sequencing issues such as flowcell defects, library degradation, and adapter read-through (Andrews, 2015).

To further investigate potential contamination, a taxonomic profile and GC% probability density was determined for each library.

The former was conducted using a custom tool dubbed “DueyDrop” which functions as follows. Briefly, for each library 5 batches of 10,000 PE reads were sampled using the reservoir sampler (Vitter, 1985) implemented in Heng Li’s seqtk library (Li, 2015). Despite 5 batches of 10,000 reads theoretically being equivalent to 50,000 random samples by splitting sampling and using a different random seed any problems from poor randomisation implementation was minimised and consistency of taxonomic profiles could be easily assessed. These randomly sampled reads were subsequently aligned to NCBI’s Protein NR RefSeq database (Pruitt et al., 2007) using the efficient short-read optimised BLASTX implementation of DIAMOND (Buchfink et al., 2015) (at a expectation of e^{-5} and top hits for each read retained. Gene identifiers (GI) were extracted from these tops hits and queried against a local copy of the NCBI taxonomy database (Federhen, 2012) to recover a hit taxonomic lineage for each read that aligned to a sequence within NR database. These lineages were then interactively tallied at several different taxonomic levels (e.g. domain level - eukaryote vs bacteria, or lower level - viridiplantae vs ciliate) and variances calculated. Results were then tabulated and libraries compared to assess whether any libraries appeared aberrant. This whole analysis was repeated for both untrimmed reads and reads quality trimmed to a high quality threshold of an average Q₃₀ over a sliding window of size 4 using Trimmomatic (Bolger et al., 2014) to assess the impact trimming has on this profiling. Taxonomic profiles were additionally visualised in Krona (Ondov et al., 2011) using the tabular BLAST hit import functionality.

Scripts used to conduct this analysis are available in the following github repository:

<https://github.com/fmaguire/dueydrop>

To determine how representative profiles created using small subsamples consisting of <1% of reads are to profiles of entire libraries a similar analysis was done using full libraries. All libraries were pre-trimmed at the harsh threshold of the Q₃₀ sliding window discussed above. The forward read from each trimmed library was used used in a similar DIAMOND based BLASTX search however all hits were retained. Multiple hits for a given read were collapsed into a single lowest common ancestor (LCA) using the LCA algorithm (Gabow and Tarjan, 1985) implemented in MEGAN (via the “mtools” package) (Huson et al., 2007; El Hadidi et al., 2013). LCA were then summarised and tabulated using a script in the CGAT collections (“lca2table.py”) (Sims et al., 2014) and visualised using Krona (Ondov et al., 2011).

On the basis of the resultant taxonomic profiles libraries were excluded or included from downstream preprocessing and assembly. The libraries selected for inclusion during these analyses are referred to as the “taxonomically filtered” single cell libraries.

4.3.2.2 GC DENSITY ESTIMATES

Each library's GC% probability density was estimated from per-read GC proportions (calculated using awk ([Aho et al., 1987](#))) via Kernel Density Estimation (KDE) ([Rosenblatt, 1956](#); [Parzen, 1962](#)) (implemented in the `seaborn` package ([Waskom et al., 2015](#))). This involved a standard gaussian kernel and a bandwidth determined by "Scott's normal reference rule" ([Scott, 1979](#)). Again this analysis was repeated with both untrimmed and Q₃₀ trimmed reads.

4.3.3 OPTIMISING READ PRE-PROCESSING

4.3.3.1 TRIMMING

To investigate the optimal trimming parameters for single cell libraries, random subsamples were trimmed using a range minimum quality thresholds and then the effects investigated by mapping against 3 draft *de novo* transcriptomes.

Specifically, 5000 PE reads were randomly sampled without replacement from each of the raw FASTQ libraries using the streaming reservoir sampling ([Vitter, 1985](#)) algorithm implemented in Heng Li's `seqtk` C library (([Li, 2015](#))). To guarantee that pairing was maintained the same random seed was used for the left and right read of each library and incremented between libraries.

Trimmomatic ([Bolger et al., 2014](#)) was run on these samples with adapter clipping (ILLUMINACLIP) using sequencing service provided fasta file of adapters, a maximum mismatch count of 2, a palindromic clip threshold of Q₃₅ and a simple clip threshold of Q₁₅, a sliding window quality trim of size 4 and average window quality thresholds of Q₀, Q₂, Q₅, Q₁₀, Q₁₅, Q₂₀, Q₂₅, Q₃₀, Q₃₅, and Q₄₀. Finally, a minimum length trimmed read length criteria of 40bp was used.

The trimmed samples were then mapped to 3 different *de novo* draft transcriptome assemblies using `bowtie2` ([Langmead and Salzberg, 2012](#)) with maximum and minimum insert sizes of 37bp and 1161bp (derived from library preparation fragment size distribution and histograms of mapped insert sizes for untrimmed reads against bulk reference).

These 3 draft assemblies were a "baseline" bulk RNASeq transcriptome reference consisting of a Trinity ([Haas et al., 2013](#)) assembly of the light and dark bulk libraries preprocessed to remove low quality bases (<Q₂₀) and adapters using Fastq-MCF ([Aronesty, 2013](#)); and two Trinity assemblies of the taxonomically filtered sc-RNASeq libraries previously trimmed at an average window quality threshold of Q₅ and Q₃₀ respectively.

For each library and set of quality thresholds the total number of concordantly mapping (i.e. forward and reverse PE reads mapped to transcripts within the range of the insert sizes used) reads was recorded. This heuristic measure was chosen because the number of concordantly mapping reads generally correlates with the assembly quality ([MacManes, 2014](#)). The proportion of surviving reads which mapped was not used as a metric because this could be spuriously inflated in cases where a particular set of trimming parameters has caused the majority of reads to be discarded.

The number of concordantly mapping reads were tallied and plotted in `seaborn` for each library, reference

transcriptome and set of trimming parameters. The shape of this line was then used to determine the optimal quality threshold to use for further assembly.

Scripts used to conduct this are available in my thesis scripts github repository: https://github.com/fmaguire/thesis_scripts/tree/master/chapter_2_assembly_and_binning/trimming_optimisation

4.3.3.2 GC PARTITIONING OF READS

To assess the utility of pre-assembly read partitioning an unsupervised clustering tool was created: Paired Arrangement of Reads via K-means On Unlabelled Reads (parKour). This C++ tool implements a fast and efficient K-means clustering of reads based on the dual features of GC% in forward and reverse paired reads and designed to exploit the wildly differing GC biases of *P. bursaria* and *M. reisseri*.

ParKour operates as follows:

1. Parse user input of paired FASTQ files corresponding to Forward and Reverse Paired-End reads, and desired number of clusters
2. Simultaneously iterate over the pair of FASTQ files calculating the GC% for each loading results into an Armadillo $2 \times N$ matrix ([Sanderson, 2010](#)) where N is the total number of PE reads
3. Bradley-Fayyad K-means ([Bradley and Bradley, 1998](#)) clustering as implemented in the MLPACK library ([Curtin et al., 2013](#))
4. Re-read the two input FASTQs assigning them to output files based on the assigned cluster of the pair

GNUpot ([Williams et al., 2010](#)) was used to visualise classification and cluster assignment. This approach was attempted using a range of expected clusters from 2 to 5.

Scripts used to conduct this are available in a github repository: <https://github.com/fmaguire/parKour>

4.3.3.3 ERROR CORRECTION

The effect of error correction on assemblies involving single cell libraries was assessed by applying two different error correction algorithms to the screened, trimmed reads before assembly. These were a Bayeshammer ([Nikolenko et al., 2013](#)) implemented as part of the Spades genome assembler ([Bankevich et al., 2012](#)) and optimised for MDA-based single cell genomic data, and “SEECER” ([Le et al., 2013](#)) which is optimised for RNAseq (but not necessarily sc-RNAseq data). The impact of each of these error correction algorithms at the read level was assessed as well as their subsequent impact on downstream assembly metrics, particularly RSEM-EVAL likelihood score as will be expanded upon below in the description of assembly assessment.

4.3.3.4 K-MER NORMALISATION AND TRIMMING

Taxonomically screened sc-RNAseq libraries trimmed at a minimum sliding window quality threshold of Q₃₀ and bulk libraries were K-mer normalised and trimmed using the Khmer package ([Crusoe et al., 2015](#))

Specifically, reads were interleaved (Döring et al., 2008) and then digitally normalised using diginorm (Brown et al., 2012) with a K-mer size and coverage cut-off of 20. Low abundance and likely erroneous K-mers were then filtered relative to the read coverage i.e. low abundance k-mers were removed from high coverage reads but would be more likely to be retained for low coverage reads (Zhang et al., 2015, 2014).

Filtered data was then assembled using Trinity (with minimum K-mer coverage of 2) and the subsequent assembly partitioned into transcript families in Khmer (Pell et al., 2012).

The final assemblies were then compared to un-normalised and k-mer trimmed assemblies (see section 4.3.4.1 for details.

4.3.4 ASSEMBLY

Referenced and *de novo* assemblies were attempted using a range of assemblers and assembly parameters.

Firstly, trimmed bulk and taxonomically filtered single cell libraries were mapped to *Chlorella NC64A*, *Coccomyxa C169*, *Tetrahymena thermophila* and *Paramecium caudatum* macronuclear (MAC) genomes. The former pair being the closest available genomes to the endosymbiont and the latter to the host. Mapping was done using the TopHat2 spliced aligner (Kim et al., 2013) against the genomes and was supplemented with and without annotated ORF information (in the form of gtf). GTF files were generated from best available gene annotations in the form of GFF files using gffread (part of cufflinks). Cufflinks (Trapnell et al., 2011) was then used to extract isoforms from the spliced alignments.

For *de novo* assembly, assemblies were conducted using following assemblers with default settings unless specified otherwise:

- Trinity v2.0.6 (Grabherr et al., 2011) with and without a minimum K-mer coverage of 2
- SOAPdenovo-Trans v1.03 (Xie et al., 2014) with K-mer sizes of 20, 32, 64, and 80
- TransAbyss v1.5.3 (Robertson et al., 2010) with K-mer sizes 20, 32, and 64
- Velvet v1.2.10 (Zerbino and Birney, 2008) and Oases v0.2.08 (Schulz et al., 2012) with K-mer size of 21, a minimum K-mer coverage of 2 and a minimum transcript length of 100.
- Iterative de Bruijn Graph Assembler (IDBA)-tran (Peng et al., 2010, 2012, 2013)
- IDBA-MTP (Leung et al., 2014), IDBA-UD (Peng et al., 2012), IDBA-MT (Leung et al., 2013) workflow.
- Bridger (Chang et al., 2015).

Trinity was used for all further downstream assembly optimisation due to its performance and consistency. Specifically, a minimum K-mer coverage of 1-3 were attempted as well as various combinations of libraries (i.e. bulk and screened sc-RNAseq libraries) and also sequencing data from Kodama's previously published *P. bursaria* bulk RNAseq analysis (Kodama and Fujishima, 2014).

To assess the utility of combining assemblies as discussed in (Nakasugi et al., 2014), the best assemblies from Bridger and Trinity (as assessed below) were combined using the EvidentialGene tr2aacds pipeline (Gilbert,

2013). Additionally, the best assemblies from all assemblers that ran to completion i.e. Bridger, Trinity, SOAPdenovo-Trans, Transabyss and IDBA-tran were also combined and assessed.

4.3.4.1 ASSEMBLY ASSESSMENT

Resultant assemblies were compared using standard assembly statistics (e.g. contigs number and size, bases assembled) as implemented in a perl script supplied with Trinity (Haas et al., 2013) and TransRate (Smith-unna et al., 2015). Additionally, reference free probabilistic assembly assessment RSEM-EVAL package (part of DETONATE) (Li et al., 2014) to produce likelihood scores for various completed assemblies.

4.3.4.2 ORF CALLING

ORFs were called from assembled transcripts using TransDecoder (Haas et al., 2013) with a minimum protein size of 100aa.

TransDecoder operates as follows

1. All ORFs are found in transcripts by identification of sequences between a start codon and an in-frame stop codon. Partial ORFs are also identified as sequences between the 5' transcript terminus and a stop codon or a start codon and the 3' transcript terminus.
2. The top 500 longest of these ORFs was selected and used to train a reading-frame specific 5th-order Markov model.
3. All of the ORFs were then scored for each reading frame as a sum of the per-base log odd scores (log probability of a given base and reading frame given its preceding 5 bases normalised by the relative frequency of that nucleotide across all transcripts).
4. The highest scoring reading frame is retained as a candidate
5. Any of the initial ORFs with homology to proteins in PFAM and Swissprot (as determined by HMMR and BLASTP (minimum e-value of $1e^{-5}$)) are also retained.

Paramecium uses an alternative genetic code in which two universal stop codons (UAA, UAG) are reassigned to glutamine. For the purposes of initial BLAST based binning ORFs were called and translated using both universal encoding and this alternative code. However, for the later BLAST-based bin accuracy verification purposes and subsequent automated phylogeny based binning all ORFs were initially only called using the alternative ciliate encoding. The ciliate encoding was used instead of universal because it was spuriously extended transcripts were considered favourable to falsely truncated ones. This greatly reduced redundancy in the later binning analyses.

4.3.5 TRANSCRIPT BINNING

4.3.5.1 INITIAL BLAST BASED BINS

Initially, 10,000 randomly chosen, translated transcripts from an earlier iteration of the assembly process were binned into their predicted source - Host (H), Endosymbiont (E), Food (F) and Unknown (U).

Each of the assembled transcripts were BLASTP-ed against a database consisting of the following predicted proteomes: *Chlorella* NC64A, *Chlamydomonas reinhardtii*, *Coccomyxa* C169, *Paramecium tetaurelia*, *Tetrahymena thermophila*, *Arabidopsis thaliana*, *Homo sapiens* (helping to identify contamination), *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Bacillus cereus* ATCC 14579, *Escherichia coli* 536, *Escherichia coli* O157 H-7, *Salmonella typhimurium* LT2 and *Escherichia coli* K-12 (the last 5 genome datasets helping to identify food bacterial genes).

Then initial bins were determined as follows:

- Endosymbiont (E): Transcript's highest scoring BLAST hit at an expectation of $\leq e^{-50}$ was to *Coccomyxa*, *Chlamydomonas* or *Chlorella*. Or transcript's highest scoring hit at e^{-20} was one of those species and the longest likely coding region in the transcript was using the universal codon table.
- Host (H): Transcript's highest hits at $\leq e^{-50}$ were to *Paramecium tetaurelia* or *Tetrahymena thermophila*. Or highest hit at e^{-20} was one of those species and longest likely coding region was using the *Tetrahymena* codon table.
- Food (F): Transcript's highest scoring BLAST hit at an expectation of $\leq e^{-50}$ was to one of the *E. coli* species or *Salmonella*. Or transcript's highest scoring hit at e^{-20} was one of those species and the longest likely coding region in the transcript was using the universal codon table.
- Unknown (U): highest scoring hits to *Arabidopsis*, *Homo sapiens*, *Saccharomyces* or *Schizosaccharomyces* or any sequence not fitting into the above categories.

The accuracy of the BLAST based binning was then determined by generating phylogenies using the method described below. Resultant phylogenies were then manually parsed and assessed for phylogenetic congruence with their bin. For example, do host binned sequences predominantly branch with other ciliate sequences? Do endosymbiont binned sequences mainly branch with archaeplastida sequences?

4.3.5.2 AUTOMATED PHYLOGENY GENERATION PIPELINE - DENDROGENOUS

To rapidly generate phylogenies an established lab tree generation pipeline, known as “Darren’s Orchard” (Richards et al., 2009) was modified and ported to python3 from perl5. This new pipeline “Dendrogenous” takes in a multi-fasta set of inputs and a set of genomes to search against. For each input sequence:

1. The user specified genome database is queried using BLASTP
2. The results are parsed and a fasta file of putative homologues is created, with inputs that have fewer than a specified number of hits (default of 5) ejected.
3. A multiple sequence alignment (MSA) is created from this fasta using Kalign (chosen for its speed) (Lassmann et al., 2009)
4. This MSA is then masked automatically to remove ambiguous sites using TrimAL (Capella-Gutiérrez et al., 2009) and masked alignments with fewer than a specified number of sites (default of 30) are ejected from the pipeline.

5. A rapid maximum-likelihood phylogenetic tree is generated using FastTree2 (Price et al., 2010)
6. Finally, encoded taxonomic information is recovered from the “cider” database of the original “Darren’s Orchard” pipeline and the trees are named with full species names.

The two key improvements are that of full and efficient parallelisation of the tree generation process (see fig. 4.3.1) and increased use of filestreams to pass data between pipeline stages. This latter modification reduces costly and slowly file reading and writing operations.

In the process of creating this modified phylogenetic pipeline I upgraded the general purpose python phylogenetic toolkit ETE (Huerta-Cepas et al., 2010) to support python3. As ETE is an open source project I submitted these changes to the maintainer and they have subsequently been merged into the master. These changes compose a significant proportion of the latest major release version of this toolkit <https://github.com/jhcepas/ete/pull/105>.

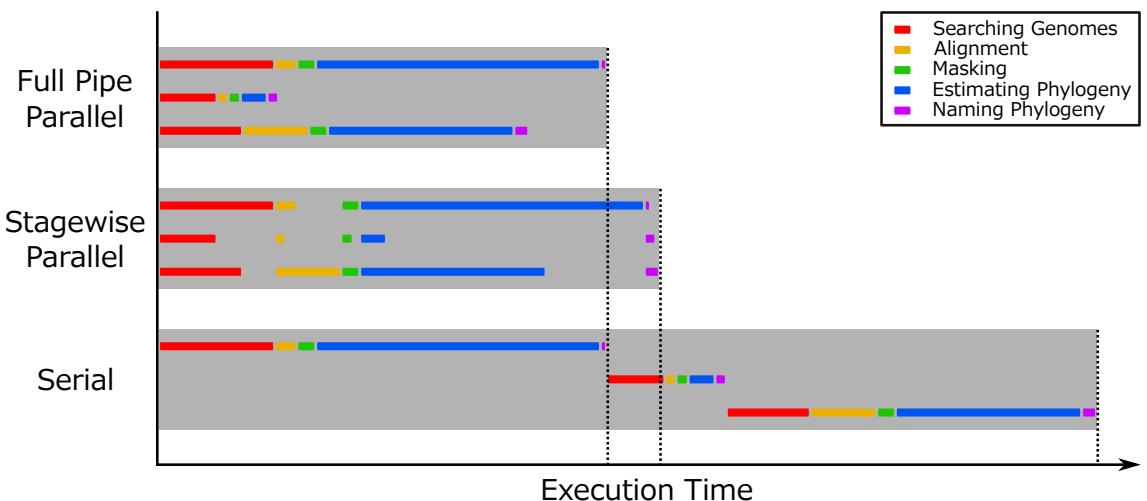


Figure 4.3.1: A explanatory plot showing 3 different possible architectures for a tree generation pipeline. Serial, in which each phylogeny is run one after another. This form makes no use of multiprocessing facilities, however, a moderate but significant performance improvement can be achieved by allowing each stage in the pipeline to utilise multiple cores i.e. the trees are generated serially but during their generation alignment and blasting making use of multiple processors. Stagewise parallel, where for example, all alignments for each input sequence are run side-by-side and masking begins once the last sequence has finished alignment. The disadvantage of this is a single slow stage for one input sequence can hold up the whole pipeline and leave resources idle. Additionally, by running many of the same type of process at the same time, each with similar resource requirements, the risk of hardware bottlenecking is increased compared to a more heterogeneous load. Finally, fully parallel runs each input sequence through the pipeline stage-by-stage separately from all other inputs to the pipeline. This prevents blocking and allows efficient using of resources.

40 genomes covering the diversity of the tree of life, with a particular focus on green algal and ciliate representatives were selected for this phylogenetic generation: *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Ostreococcus tauri*, *Micromonas pusilla* CCMP1545, *Chlorella variabilis* NC64A *Chlorella vulgaris* C-169, *Physcomitrella patens*, *Saccharomyces cerevisiae* S288C, *Neurospora crassa* OR74A, *Homo sapiens*, *Mus musculus*, *Dictyostelium discoideum*, *Paramecium caudatum*, *Paramecium tetraurelia*, *Tetrahymena thermophila* macronucleus, *Oxytricha trifallax*, *Toxoplasma gondii*, *Guillardia theta*, *Bigelowiella natans*, *Emiliania huxleyi* CCMP1516, *Aureococcus anophagefferens*, *Ectocarpus siliculosus*, *Schizosaccharomyces pombe*, *Bacillus cereus* ATCC 14579, *Escherichia coli* str. K-12 substr. MG1655, *Escherichia coli* O157 H7 str. Sakai, *Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18, Amy-

colatopsis mediterranei U32, *Aquifex aeolicus* VF5, *Borrelia burgdorferi* B31, *Chlamydophila pneumoniae* CWL029, *Chlorobium tepidum* TLS, *Deinococcus radiodurans* R2, *Caulobacter crescentus* CB15, *Sulfolobus islandicus* M.14.2S, *Nanoarchaeum equitans* Kin4-M, *Haloferax mediterranei* ATCC 33500, *Methanococcus maripaludis* S2, *Cenarchaeum symbiosum* A.

4.3.5.3 AUTOMATED PHYLOGENETIC TRANSCRIPT BINNING - ARBORETUM

In order to automate phylogeny based transcript binning the 10,000 manually verified phylogenetic bins from the initial BLAST based binning and analysis were used as a training dataset for supervised classification.

The supervised classification was implemented in a script called “Arboretum” The cardinalities of each label in training set was relatively balanced (i.e. all within the same order of magnitude) 1975 endosymbiont phylogenies, 2600 host, 3456 food, and 1969 unknown.

“Arboretum” parses phylogenies and identifies the k (default of 10) nearest branches to the seed transcript the phylogeny was generated from. The species of these closest leaves is queried taxonomically using the NCBI taxonomy local database implemented in the ETE toolkit. With a set of look-up filters e.g. sequences from ciliates can be defined as “host-like”, a set of vectors is created for each phylogeny. These are N-dimensional vectors where N is the number of class labels being used. For example, in this specific case: “endosymbiont”, “host”, “food/bacterial”, and “unknown”. The magnitude of each dimension is the summed reciprocal phylogenetic distance between the root node and all of the nearest branches that have been identified as being indicative of a certain class. Specifically, if $\gamma(x, y)$ is the phylogenetic distance between the terminal nodes x and y , $\psi(x)$ represents the look-up filters and returns the label of the terminal node x e.g. “host”. and δ_{ij} is the Kronecker delta², then for a phylogeny A :

$$X_{A, \text{class}} = \sum_{k=1}^K \left(\frac{1}{\gamma(A_k, A_{\text{root}})} * \delta_{\psi(A_k), \text{class}} \right)$$

Where in this specific example “class” is one of the set $\{\text{endosymbiont}, \text{host}, \text{food}, \text{unknown}\}$ although naturally the classes will be encoded using integer labels. Therefore, the dimensions of X are $|A|, |\text{class}|$ i.e. the number of training phylogenies by the number of pre-defined class labels.

Training data was visualised using Radial Visualisation (RadViz) (Hoffman et al., 1997; Fayyad et al., 2001). RadViz is a form of radial co-ordinate visualisation that non-linearly maps a set of N-dimensional points onto a plane for easy 2D visualisation. This mapping operates on the physical principle of “springs” anchored evenly around a unit circle with “spring” stiffness determined by the normalised $0 - 1$ value of that dimension for that point. Each point therefore rests at the point of mechanical equilibrium between the “springs” (Novakova, Lenka and Stepankova, 2006).

1,000 vectors from this training set were held out to form the test set and all models were then trained using 5-fold cross-validation (CV) on the remaining 9,000 training vectors. We evaluated Support Vector Machines

² $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

(SVMs) with both linear and radial basis function (RBF) kernels (Vapnik and Lerner, 1963), naive Bayes, K-neighbours, Decision Trees (DT) (Quinlan, 1986), DTs ensembles in a Random Forests (Breiman, 2001) and Extremely Randomised Trees (ExtraTrees) (Geurts et al., 2006), adaptively boost (AdaBoost) DTs (Freund and Schapire, 1997), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Models were trained and hyperparameters were optimised using a randomised search instead of the less efficient grid search (Bergstra and Bengio, 2012) using Bayesian optimisation in the HPOlib library (Eggensperger et al., 2013; Komer et al., 2014) over the CV-folds. Finally, each model was assessed using the held out test set and performance was evaluated by inspection of label-wise classification reports containing various metrics e.g. label F1-scores and confusion matrices.

The best performing model and hyperparameters were then used to classify the remaining unlabelled phylogenies.

4.3.5.4 TAXAASSIGN COMPARISON

To assess the performance of supervised learning and phylogeny based system (Arboretum) described above a stand-alone sequence identity binning tool TAXAssign (<https://github.com/umerijaz/TAXAassign>) was run against the 70,605 CDS sequences.

TAXAssign queried each CDS against the entire NCBI nt database. The nt BLAST database was downloaded using update_blastdb.pl script http://www.ncbi.nlm.nih.gov/blast/docs/update_blastdb.pl and TAXAssign ran BLASTN in parallel (using GNU parallel (Tange, 2011)) with a maximum of 10 reference matches per CDS a minimum necessary percentage identity for assignment to a given taxonomic level of 60, 70, 80, 95, 95, and 97 for Phylum, Class, Order, Family, Genus and Species respectively.

Results were then tabulated and compared with the Dendrogenous-Arboretum assignments.

4.4 RESULTS

4.4.1 LIBRARY CONTAMINATION SCREENING

Libraries were screened for inclusion in assemblies by inspection of their taxonomic profiles (see table 4.4.1 and table 4.4.2) as determined by DueyDrop and their GC% probability densities (via KDE).

The GC density estimates of the single cell libraries show a clear bimodal GC density with a high 70GC% peak (fig. 4.4.1) in all dark single cell libraries. With the exception of Dark1-2 and Dark2-3 this high GC peak is a greater density than the expected peak 30-50GC% (from known GC% found in genomes of sequenced relatives of both host and endosymbiont).

When these KDE are compared to the densities estimated from the Q₂₀ trimmed bulk reads (bottom right pane in fig. 4.4.1) and raw bulk RNAseq reads from (Kodama et al., 2014) (see fig. 4.4.2) it is apparent that this high GC% peak is likely originating from a high GC% bacterial contaminant in the Dark single cell libraries.

One other observation when comparing the bulk RNAseq analyses to the single cell libraries is that the main

Kernel Density Estimates of Read GC Proportion

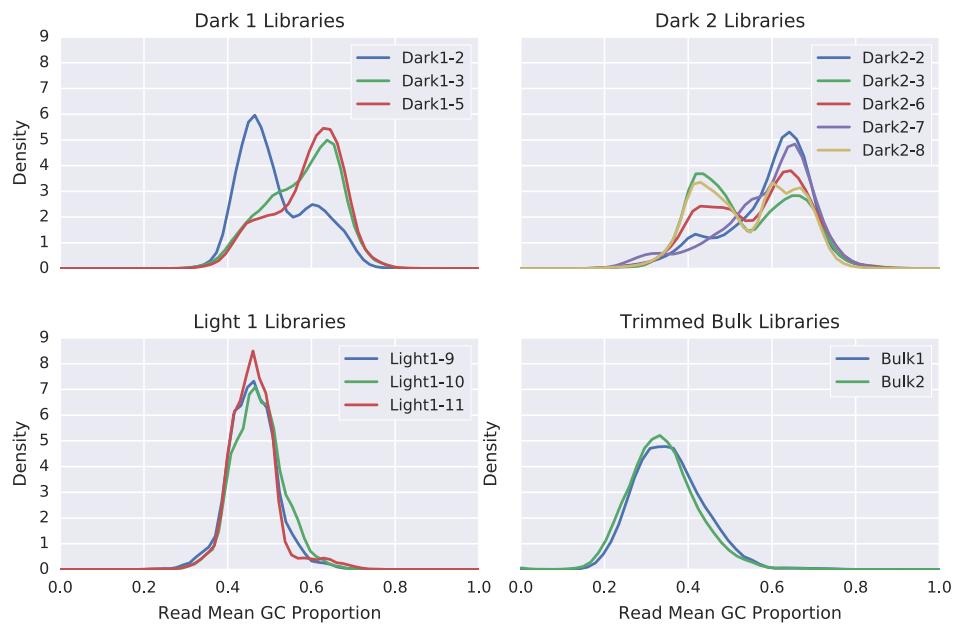


Figure 4.4.1: Probability densities of per-read GC proportions for the raw data (apart from pre-trimmed bulk explained previously) from each sequenced library. Densities were derived using Kernel Density Estimation implemented in Seaborn. Dark 1 (Dark1-2, Dark1-3, and Dark1-5) and Light 1 (Light1-9, Light1-10, Light1-11) were sc-RNAseq from the first round of SCTs sampled during the mid-dark and light culture phases. Similarly, Dark 2 (Dark2-2, Dark2-3, Dark2-6, Dark2-7, Dark2-8) were the libraries sampled in the dark from the second round of SCT. Bulk1 and Bulk2 are the bulk RNAseq libraries sequenced under lit and dark conditions. The bulk and single cell light libraries demonstrate similar shaped distributions although the bulk has a greater proportion of low GC% reads potentially representing more *Paramecium* derived data. All single cell dark libraries demonstrate a bimodal density with up to the majority of reads deriving from an unknown high 70% GC population. The dark single cell libraries exhibiting a relatively larger peak at 70% GC than at 40-50%GC (i.e. Dark1-3, Dark1-5, Dark2-2, Dark2-7) were the same libraries which were identified as potentially contaminated in taxonomic screening (see table 4.4.1).

GC peak is slightly lower in the bulk (and Kodama dataset), around 30GC% versus 45-50GC%. This possibly indicates a greater proportion of reads deriving from the low GC% *Paramecium bursaria* host and fewer from the 50GC% endosymbiont in bulk libraries relative to single cell libraries.

By comparing the results of the KDE GC analysis with and without read trimming it is apparent that trimming of reads makes nearly no difference in the density estimates. The KDE of Q₃₀ sliding window trimmed single cell reads in fig. 4.4.3 is nearly identical to that of the raw reads fig. 4.4.1.

The taxonomic profiles of single cell (table 4.4.1) and bulk libraries (table 4.4.2) generated by DueyDrop are summarised in the tables below. It is readily apparent that Dark1-3, Dark1-5, Dark2-2, and Dark2-7 display an aberrantly low number of reads aligning to known alveolate (or even eukaryote) sequences. Forward and reverse reads within a library display similar profiles with a slightly lower proportion of hits in the reverse reads. This can likely be attributed to the lower read quality found in reverse reads relative to forward reads in paired-end Illumina sequencing.

GC Proportion KDE of Kodama Read

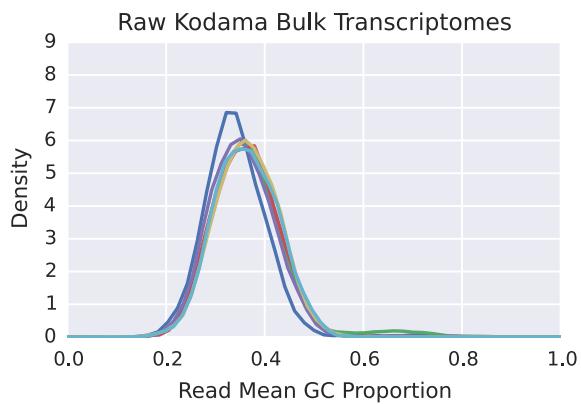


Figure 4.4.2: Probability density of the per-read GC proportion for 6 raw libraries derived from (Kodama et al., 2014) transcriptome analysis of a different *P. bursaria* species (Yad1g) with and without its *Chlorella variabilis* 1N endosymbiont. Individual libraries are indicated in the key using their DDBJ accession. This dataset displays densities relatively similar to the bulk RNAseq conducted in this project - “Trimmed Bulk Libraries” in fig. 4.4.1.

GC Proportion KDE of Trimmed Read

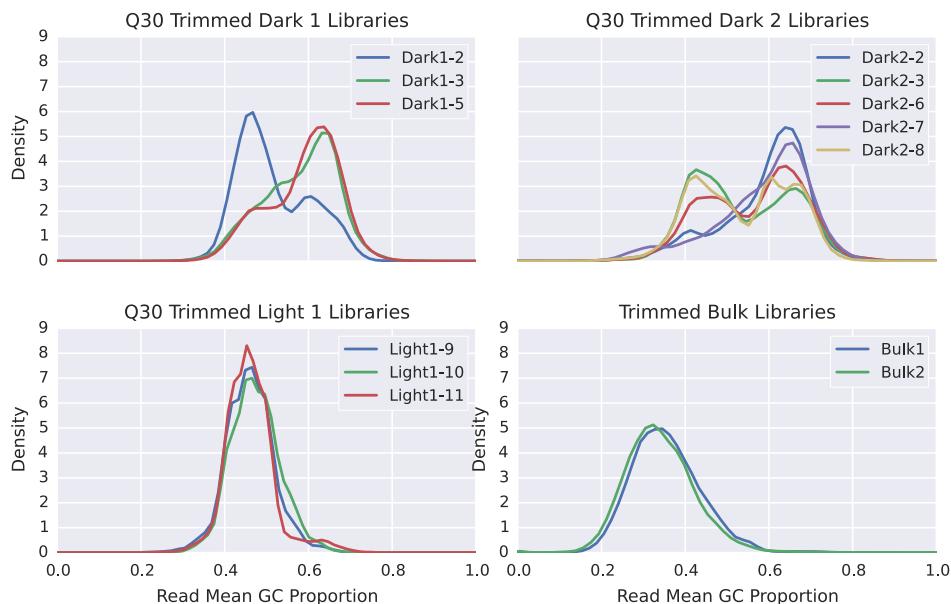


Figure 4.4.3: Probability densities of per-read GC proportions for trimmed reads. To ensure probability densities estimated in fig. 4.4.1 weren't biased by low quality ambiguous reads the same analysis was repeated using reads trimmed using a sliding window approach with a stringent average quality threshold of Q30. In all cases the densities produced appear near identical to the analysis of the raw data.

The bulk libraries demonstrate a very low level of hits compared to single cell libraries (see table 4.4.2), to the point where if they were single cell libraries they would be taxonomically excluded. However, it should be noted that the bulk libraries were sequenced on a Gene Analyzer II and are on average half the length of single cell reads (76bp vs 150bp). Due to the difficulty aligning short reads to references the difference between libraries may be attributable to this alone. Additionally, the vast majority of the lower number of hits do align to eukaryote (and

| SCT Library | PE | Eukaryote | Bacteria | Alveolate | Viridiplantae | Total Hits |
|-----------------------|----|----------------|----------------|----------------|---------------|----------------|
| <i>Light1-9</i> | R1 | 51.89 +/- 0.45 | 9.37 +/- 0.26 | 25.15 +/- 0.71 | 7.45 +/- 0.33 | 69.49 +/- 0.37 |
| | R2 | 51.75 +/- 0.25 | 8.82 +/- 0.24 | 24.85 +/- 0.56 | 7.49 +/- 0.21 | 68.75 +/- 0.29 |
| <i>Light1-10</i> | R1 | 46.35 +/- 0.56 | 15.72 +/- 0.46 | 22.96 +/- 0.24 | 6.94 +/- 0.26 | 68.73 +/- 0.30 |
| | R2 | 46.12 +/- 0.83 | 15.14 +/- 0.48 | 23.13 +/- 0.38 | 6.99 +/- 0.37 | 68.73 +/- 0.30 |
| <i>Light1-11</i> | R1 | 58.28 +/- 0.47 | 3.62 +/- 0.12 | 28.68 +/- 0.43 | 8.20 +/- 0.40 | 71.38 +/- 0.49 |
| | R2 | 57.74 +/- 0.27 | 3.50 +/- 0.10 | 28.23 +/- 0.36 | 8.41 +/- 0.31 | 70.42 +/- 0.20 |
| <i>Dark1-2</i> | R1 | 28.64 +/- 0.51 | 22.88 +/- 0.61 | 12.23 +/- 0.28 | 4.93 +/- 0.19 | 60.31 +/- 0.49 |
| | R2 | 28.29 +/- 0.24 | 21.06 +/- 0.21 | 12.13 +/- 0.28 | 4.87 +/- 0.34 | 57.65 +/- 0.35 |
| <i>Dark1-3</i> | R1 | 9.48 +/- 0.43 | 25.07 +/- 0.42 | 2.15 +/- 0.13 | 2.60 +/- 0.27 | 41.43 +/- 0.68 |
| | R2 | 8.89 +/- 0.19 | 23.11 +/- 0.52 | 2.13 +/- 0.16 | 2.45 +/- 0.18 | 38.50 +/- 0.46 |
| <i>Dark1-5</i> | R1 | 5.56 +/- 0.19 | 23.99 +/- 0.44 | 1.07 +/- 0.07 | 2.89 +/- 0.11 | 36.72 +/- 0.33 |
| | R2 | 4.94 +/- 0.21 | 21.75 +/- 0.53 | 1.02 +/- 0.11 | 2.33 +/- 0.17 | 33.06 +/- 0.52 |
| <i>Dark2-2</i> | R1 | 12.32 +/- 0.25 | 9.81 +/- 0.19 | 3.73 +/- 0.16 | 4.33 +/- 0.17 | 27.65 +/- 0.47 |
| | R2 | 11.53 +/- 0.15 | 9.00 +/- 0.17 | 3.67 +/- 0.22 | 3.74 +/- 0.12 | 25.71 +/- 0.39 |
| <i>Dark2-3</i> | R1 | 32.07 +/- 0.31 | 7.43 +/- 0.15 | 12.81 +/- 0.21 | 4.71 +/- 0.21 | 48.42 +/- 0.53 |
| | R2 | 32.47 +/- 0.24 | 6.68 +/- 0.21 | 13.11 +/- 0.43 | 4.58 +/- 0.12 | 47.92 +/- 0.28 |
| <i>Dark2-6</i> | R1 | 24.11 +/- 0.28 | 8.55 +/- 0.11 | 9.04 +/- 0.35 | 5.27 +/- 0.15 | 41.69 +/- 0.45 |
| | R2 | 22.89 +/- 0.55 | 7.44 +/- 0.17 | 8.74 +/- 0.49 | 4.36 +/- 0.24 | 38.85 +/- 0.58 |
| <i>Dark2-7</i> | R1 | 9.96 +/- 0.24 | 16.89 +/- 0.27 | 4.22 +/- 0.24 | 2.83 +/- 0.17 | 37.06 +/- 0.40 |
| | R2 | 8.77 +/- 0.18 | 15.00 +/- 0.43 | 3.94 +/- 0.14 | 2.16 +/- 0.11 | 32.86 +/- 0.29 |
| <i>Dark2-8</i> | R1 | 28.24 +/- 0.48 | 4.45 +/- 0.13 | 12.00 +/- 0.32 | 4.69 +/- 0.06 | 40.50 +/- 0.37 |
| | R2 | 28.22 +/- 0.47 | 4.30 +/- 0.22 | 11.98 +/- 0.37 | 4.32 +/- 0.24 | 40.05 +/- 0.22 |

Table 4.4.1: Taxonomic profiles of raw single cell libraries generated using “DueyDrop”. All values are percentage of reads mapping to that category +/- the standard deviation between sample replicates. The analysis was conducted for both forward and reverse reads from each library (indicated as R1 and R2 in the paired-end (PE) column). Libraries highlighted in bold were those excluded from subsequent analysis on the basis of their very low numbers of reads identifiable as eukaryotic (or specifically alveolate or archaeplastida). All forward and reverse read pairs display similar profiles to one another suggesting the problem of “MDA chimeras” may be minor.

| Bulk Library | PE | Eukaryote | Bacteria | Alveolate | Viridiplantae | Total Hits |
|--------------|----|---------------|---------------|---------------|---------------|----------------|
| <i>Light</i> | R1 | 9.66 +/- 1.55 | 0.18 +/- 0.13 | 6.28 +/- 1.41 | 0.86 +/- 0.3 | 10.10 +/- 1.48 |
| | R2 | 9.62 +/- 0.81 | 0.26 +/- 0.09 | 6.58 +/- 0.36 | 1.04 +/- 0.41 | 10.16 +/- 0.95 |
| <i>Dark</i> | R1 | 4.90 +/- 0.78 | 0.36 +/- 0.11 | 3.14 +/- 0.58 | 0.50 +/- 0.16 | 5.40 +/- 0.93 |
| | R2 | 5.50 +/- 1.25 | 0.22 +/- 0.19 | 3.82 +/- 0.81 | 0.50 +/- 0.12 | 6.02 +/- 1.22 |

Table 4.4.2: Taxonomic profile of the two trimmed (Q20) bulk transcriptome libraries generated using “DueyDrop”. All values are the percentage of reads mapping to that taxonomic category +/- the standard deviation between sampling replicates. The analysis was conducted for both forward and reverse reads from each library (indicated as R1 and R2 in the paired-end (PE) column). Overall only a very small number of bulk reads could be assigned to any taxonomic class by “DueyDrop”.

alveolate) taxa consistent with a non-contaminated library.

To identify the likely source of the high GC% contamination and to assess how representative the taxonomic profiling of small random subsamples of reads $\leq 1\%$ to full scale analyses Krona was used to create interactive hierarchical plots of the taxonomic profiles³. From this, Rhizobia species are the most prevalent high GC% species found in the libraries with this high 70GC% peak in the KDE plots and therefore are the most likely source of this particular aspect of contamination.

Therefore, small random subsamples are representative of the full library and read-level taxonomic assignment can be used to screen single cell libraries for contamination.

³Accessible at http://finlaymagui.re/dueydrop_analysis

Taxonomic Profiles of
Representative Excluded Libraries

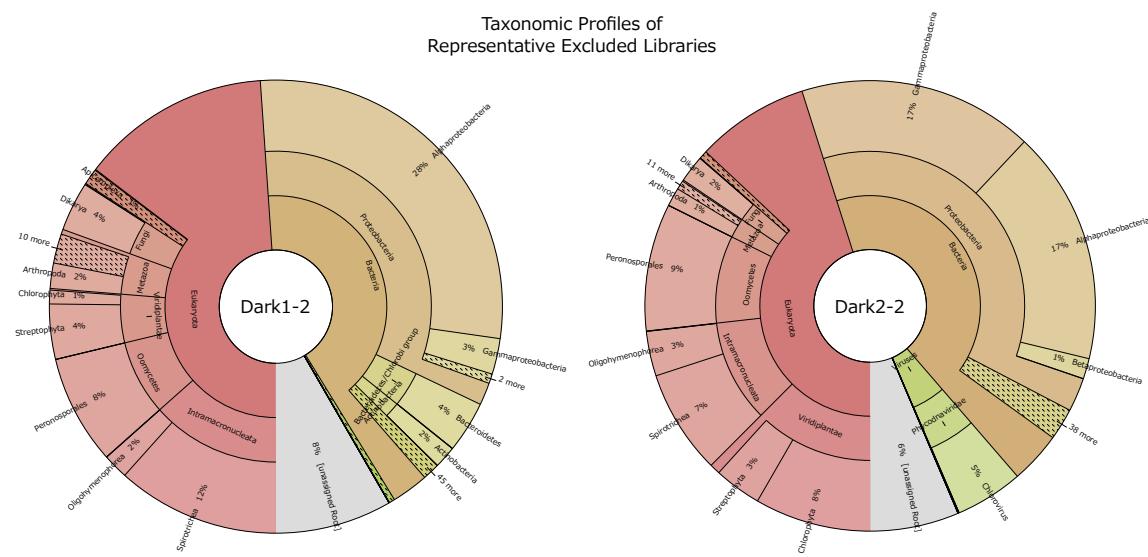


Figure 4.4.4: Krona visualisation of taxonomic profiles of two representative single cell libraries (Dark1-2, Dark2-2) that were excluded from further analysis due to aberrant profiles (typically large proportion of reads being assigned to Bacteria than Eukaryota). Note that nearly 50% of each library is identified as bacterial.

Taxonomic Profiles of
Representative Included Libraries

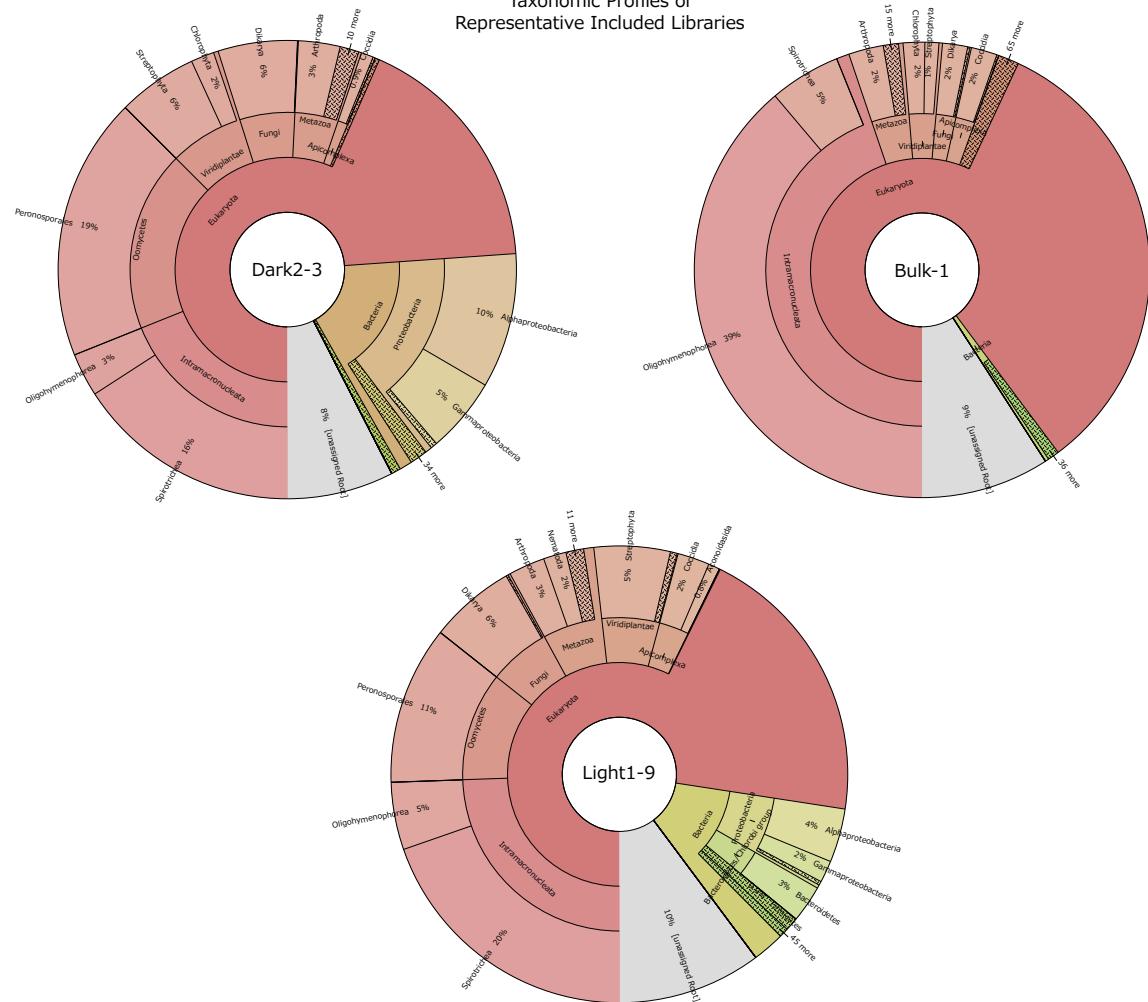


Figure 4.4.5: Krona visualisation of the taxonomic profiles of representative RNAseq libraries (Bulk1, Dark2-3, and Light1-9) that were retained in the analysis after taxonomic screening. The key thing this figure shows is that in retained libraries the vast majority of reads were identified as eukaryotic in origin.

Taxonomic Profiles from Subsamples Compared To Full Analyses

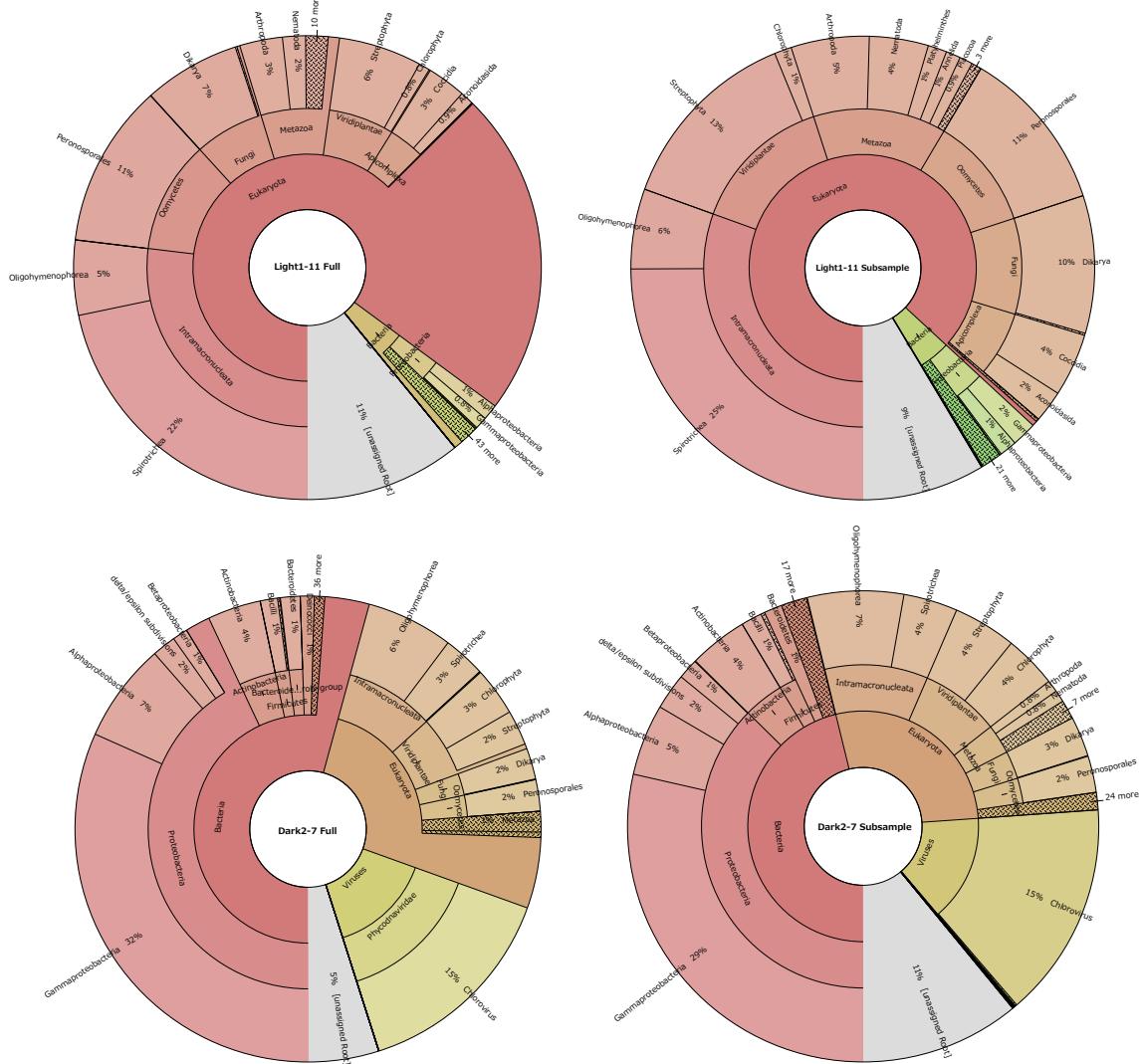


Figure 4.4.6: Comparison of taxonomic profiles derived from small $\leq 1\%$ random subsamples of libraries compared to profiles generated using the full library. Light1-11 and Dark2-5 are used as representative examples as they display the trends common for all single cell libraries. All subsamples demonstrated taxonomic profiles with relatively similar proportions to full analyses. For example, in the Light1-11 subsample of reads with hits the proportion of eukaryote to Bacteria was 87:4 % vs 85:4% of the root for the full analysis. Similarly the ratios for Dark2-7 shown eukaryote to bacteria are 26:54 for full analysis and 28:46 for subsample. The key difference is the assignment of a greater proportion of reads to intermediate taxonomic levels in the full analyses due to the difference in resolution of multiple hits per read. Principally, the full library analyses retain all hits and assign level based on a lowest common ancestor algorithm whereas the subsample analysis just uses the top hit.

4.4.2 READ PRE-PROCESSING

4.4.2.1 TRIMMING OPTIMISATION

The optimal trimming threshold was determined by a combination of read mapping statistics against 3 preliminary reference assemblies as well as the impact on resultant *de novo* assemblies at that threshold.

A rapid decrease in the number of concordantly mapping PE reads (i.e. within insert distance of one another) was observed above a Q₃₀ quality threshold. This proves true regardless of the reference assembly being mapped to (see fig. 4.4.7). Q₃₀ relative to Q₂₀ appears to induce a very slight decrease in total number of mapping reads but not drastically so.

Additionally, naive assemblies in Trinity of taxonomically screened single cell libraries at different sliding window quality threshold trims of Q₅, Q₂₀, and Q₃₀ (table 4.4.3) were created. These show that more permissive trims (Q₅ and Q₂₀) lead to a greater number of assembled bases and transcripts however the likelihood of these assemblies are also lower than that generated using the more conservative Q₃₀ trim. However, it should be noted that the difference in the number and size of assembled transcripts at different thresholds was less than was found using different assemblers and assembly parameters.

| Trim Threshold | Number of Transcripts | Bases Assembled | Assembly Likelihood (- log) |
|-----------------|-----------------------|-----------------|-----------------------------|
| Q ₅ | 112,182 | 52,511,552 | -3.168 * 10 ¹⁰ |
| Q ₂₀ | 107,955 | 50,809,686 | -3.015 * 10 ¹⁰ |
| Q ₃₀ | 99,784 | 47,313,963 | -2.832 * 10 ¹⁰ |

Table 4.4.3: Comparison of Trinity assemblies of taxonomically screened single cells reads (no bulk reads) at 3 different sliding window minimum average quality trimming thresholds. Trimming largely does not cause a major difference between assemblies in terms of number of contigs recovered or overall assembly likelihoods. Harsher (Q₃₀) trims result in slightly smaller but slightly more likely assemblies than permissive trims (Q₅).

Therefore, due to increasing the assembly likelihood while only very marginally decreasing the number of contigs and mapping reads relative to more permissive trims Q₃₀ was determined to be the optimal trimming threshold. It can be considered from this data that Q₃₀ forms a maximum feasible stringency for trimming.

| Library | Number of raw PE Reads | Number of Q ₃₀ trimmed PE Reads |
|-----------|-------------------------|--|
| Dark1-2 | 6.460 * 10 ⁷ | 3.355 * 10 ⁶ |
| Dark2-3 | 2.243 * 10 ⁷ | 1.478 * 10 ⁷ |
| Dark2-6 | 2.431 * 10 ⁷ | 1.443 * 10 ⁷ |
| Dark2-8 | 2.761 * 10 ⁷ | 1.866 * 10 ⁷ |
| Light1-9 | 1.524 * 10 ⁷ | 1.382 * 10 ⁷ |
| Light1-10 | 1.614 * 10 ⁷ | 1.478 * 10 ⁷ |
| Light1-11 | 1.474 * 10 ⁷ | 1.334 * 10 ⁷ |

Table 4.4.4: Summary of the library size of the taxonomically selected single cell libraries before and after trimming at a minimum average SLIDINGWINDOW quality threshold of Q₃₀. Of interest, Dark1-2 was generally of poor quality and thus was disproportionately minimised by trimming. Additionally, the two bulk RNAseq libraries were trimmed at Q₂₀ in FastQ-MCF resulting in total library sizes of 2.458 * 10⁷ and 2.779 * 10⁷ respectively

Comparison of Trimming Parameters on Mapping

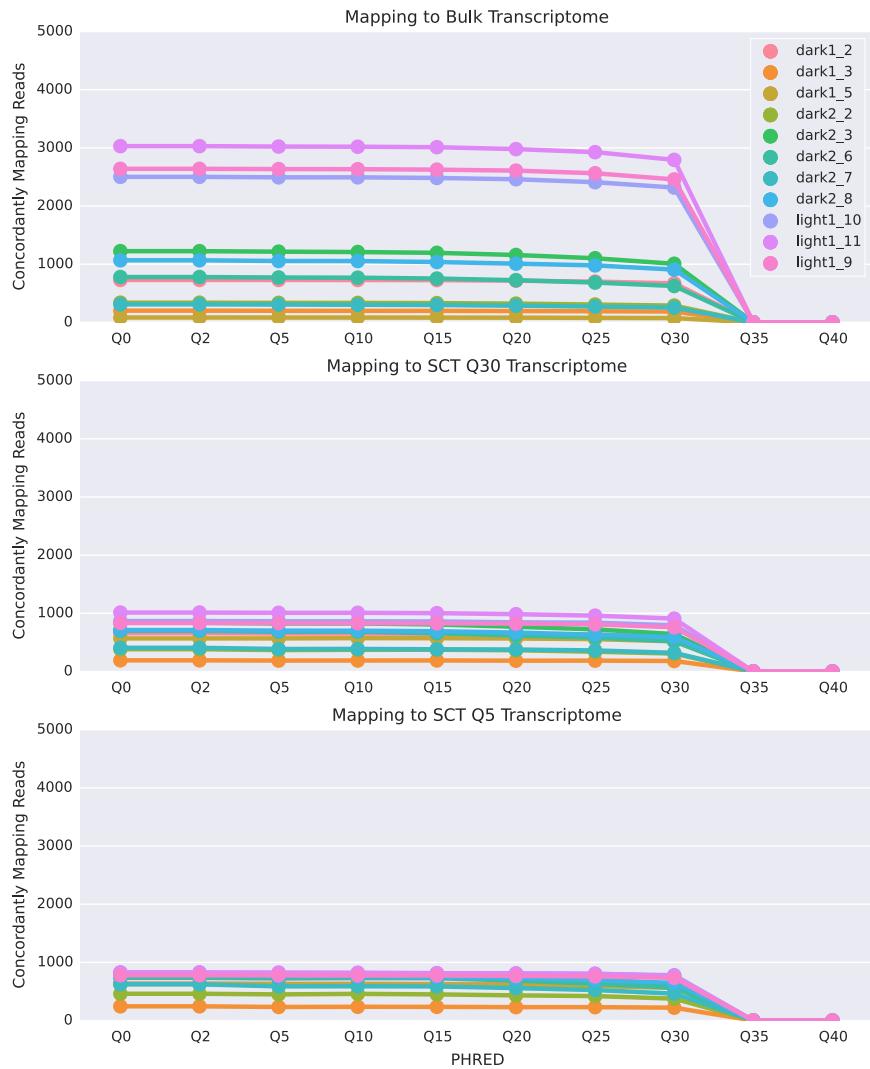


Figure 4.4.7: Assessment of the optimal minimum average quality threshold in Trimmomatic's sliding window (size 4) trim. Plots display the number of concordantly mapping reads (i.e. the forward and reverse read map to assembly at a distance of approximately their insert) at a range of different trimming thresholds. 5000 randomly sampled PE reads from each single cell library are mapped against 3 different reference assemblies. The key finding is above a threshold of Q30 there is a huge decrease in the number of mapping reads.

| Clustering Scheme | Centroids | Number of Reads Assigned |
|---------------------|------------------|--------------------------|
| 2 | (0.6674, 0.6177) | 57.3M |
| | (0.4557, 0.4393) | 81.6M |
| 2 (over-clustering) | (0.6672, 0.6168) | 57.7M |
| | (0.4555, 0.4392) | 81.2M |
| 3 | (0.5363, 0.5092) | 44.0M |
| | (0.6924, 0.6394) | 43.3M |
| | (0.4231, 0.4096) | 51.6M |
| 3 (over-clustering) | (0.5365, 0.5090) | 43.9M |
| | (0.6921, 0.6396) | 43.6M |
| | (0.4235, 0.4098) | 51.7M |

Table 4.4.5: Final cluster centroids and number of reads assigned to each cluster in parKour using various run settings. Centroids are the mid-point of each cluster, therefore in the 2 cluster scheme “parKour” identified one cluster of reads centred around 66.74% GC for the forward read and 61.77% for the reverse read. Note that overclustering made a minimal impact on cluster location.

4.4.2.2 GC PARTITIONING

GC partitioning was conducted on Q₃₀ trimmed reads using K-means clustering as implemented in the parKour tool described above to attempt to remove GC% rich contamination from single cell libraries.

The 2 different clustering schemes attempted using 2 and 3 target clusters. Additionally, both clustering schemes were also run with an initial overclustering factor of 3 i.e. parKour originally found 6 and 12 clusters and then merged them to produce the target 2 and 3 clusters respectively.

2 and 3 target clusters with and without an initial overclustering factor of 3 (i.e. initially finding 6 and 12 clusters originally before merging to produce final 2 and 3 cluster targets).

Therefore, over-clustering made a minimal effect on cluster centroids and read assignment.

Unfortunately, as might have been foreseen, the resultant assemblies from individual read clusters displayed high levels of fragmentation regardless of the clustering regime used. For example, in the case of the 2 cluster (without over-clustering) and subsequent individual Trinity-based assemblies resulted in 268,806 transcripts of marginally shorter average length than the equivalent un-pre-partitioned assembly (99,784 transcripts).

The same pattern, consistent with assembly fragmentation, was observed when only dark single cell libraries were clustered using 2 or 3 clusters. Therefore, GC-based pre-assembly read partitioning proved incapable of improving the assembly of this highly heterogeneous RNAseq dataset.

4.4.2.3 ERROR CORRECTION

Error correction was attempted on both lightly trimmed (Q₅) and harshly trimmed (Q₃₀) taxonomically selected SCT reads.

Bayeshammer, as implemented in the Spades genome assembler, even on permissively trimmed ($Q > 5$) reads corrected only a maximum 0.0007% of reads in the 7 taxonomically selected SCT libraries. As this affected on the order of 10s of reads it was not considered worth pursuing this tool further.

“SEECER”, an RNA-Seq specific error correction tool was used to correct lightly trimmed (Q₅) and harshly trimmed (Q₃₀) SCT reads. Approximately, 5.37% of Q₅ trimmed SCT reads were corrected in “SEECER”. 0.51%

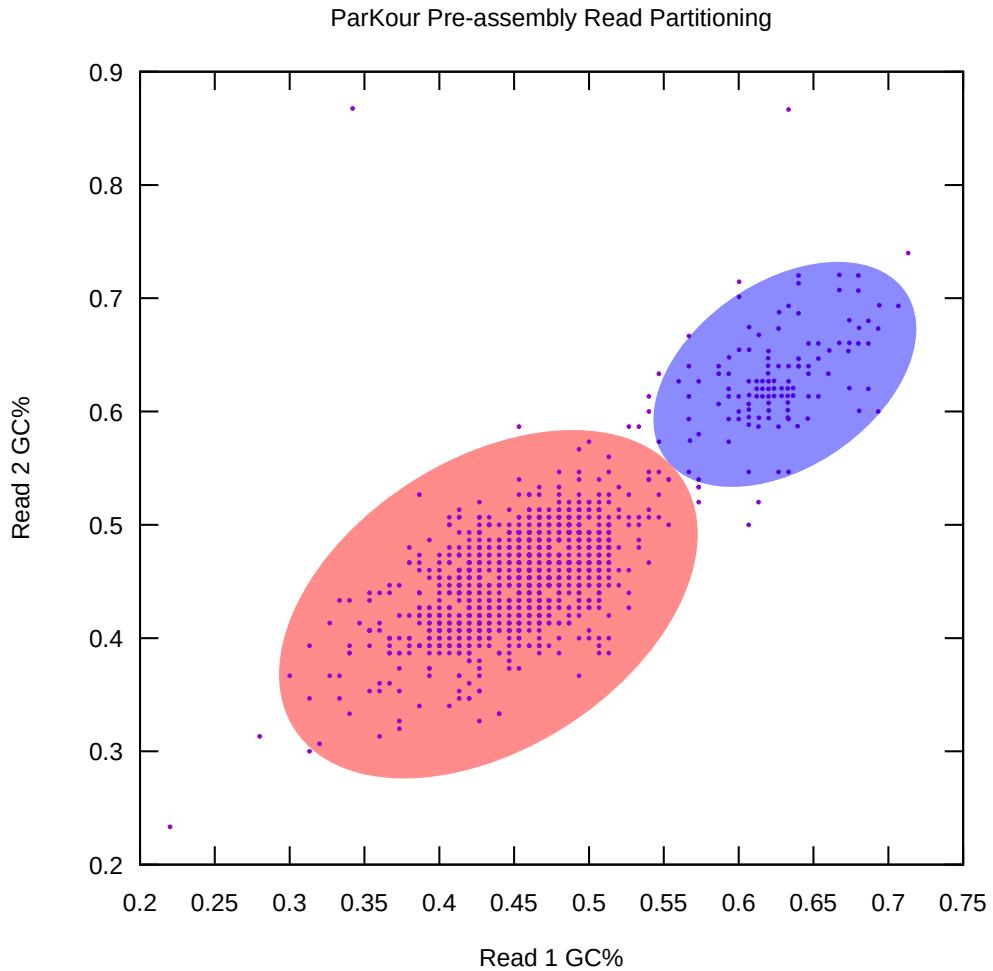


Figure 4.4.8: Visualisation of GC-based paired read K-means clustering on a small random subset of all single cell transcriptome reads. 2 initial centroids were specified without an overclustering factor and approximate final centroids (0.6674, 06177) and (0.4557, 0.4393) are indicated by highlighted areas. 57.3M and 81.7M were assigned to each respective cluster. Comparison to other clustering regimes can be found in table 4.4.5. This supports the finding of the fig. 4.4.1 that there is a clear and identifiable cluster of high GC% reads present in the sample and it is possible to identify and group these reads using unsupervised learning.

of Q₃₀ trimmed SCT reads were corrected.

Trinity assemblies of taxonomically selected single cell libraries (without bulk libraries) were then compared with and without “SEECER” error correction (see table 4.4.6).

| Trim Threshold | Number of Transcripts | Bases Assembled | Assembly Likelihood (- log) |
|----------------------------------|-----------------------|-----------------|-----------------------------|
| Q ₅ | 112,182 | 52,511,552 | $-3.168 \cdot 10^{10}$ |
| Q ₅ SEECER Corrected | 111,853 | 51,847,128 | $-3.147 \cdot 10^{10}$ |
| Q ₃₀ | 99,784 | 47,313,963 | $-2.912 \cdot 10^{10}$ |
| Q ₃₀ SEECER Corrected | 96,494 | 46,312,469 | $-2.995 \cdot 10^{10}$ |

Table 4.4.6: Naive Trinity assembly of Q5 and Q30 trimmed taxonomically selected single cell libraries with and without SEECER error correction. While assembly likelihood increases after error correction for Q5 trimmed reads it is still lower than Q30 uncorrected. For Q30 trimmed reads error correction marginally decreases assembly likelihood.

As can be observed, error correction of SCT reads made minimal effect in the overall likelihood of assemblies for this dataset even with lightly trimmed reads. Error corrected Q₅ trimmed reads performed worse than

Q_{30} trimmed reads without error correction. Additionally, Q_{30} trimmed reads generated marginally less likely assemblies with error correction than without.

Therefore, error correction was considered ineffective for this dataset and thus was not used for further analysis. Instead, we elected to use uncorrected, taxonomically selected, Q_{30} trimmed reads from this point on.

4.4.2.4 DIGITAL NORMALISATION

Digital normalisation and removal of likely erroneous k-mers (i.e. low abundance) via Khmer reduced the total input reads from the Q_{30} trimmed taxonomically filtered SCT and bulk libraries from $2.912 \cdot 10^8$ to $8.473 \cdot 10^6$ paired reads.

Of those, $6.231 \cdot 10^6$ derive from the bulk and $2.253 \cdot 10^6$ from single cell libraries. Therefore, as Q_{30} trimmed single cell libraries comprised $9.318 \cdot 10^6$ paired end reads and bulk libraries consisted of $52.377 \cdot 10^6$ reads digital normalisation and abundance filtering resulted in a retention of 2.418% of single cell PE reads and 11.891% of bulk PE reads.

Of these surviving single cell PE reads $9.762 \cdot 10^5$ were from the 3 selected light libraries (Light1-9, Light1-10, and Light1-11) and $1.277 \cdot 10^6$ were derived from the dark libraries (Dark1-2, Dark2-3, Dark2-6, and Dark2-8). Therefore, abundance filtering and digital normalisation did not disproportionately remove light or dark single cell reads.

This Khmer based pre-processing had a very positive effect on assembly likelihoods. The standard Trinity assembly improved in likelihood by an order of magnitude while assembling more transcripts of near equal length (based on median contig length). The Khmer processed assembly marginally increased median contig length at the expense of a lower N₅₀.

| Preprocessing | Number of Transcripts | Bases Assembled | Contig N ₅₀ | Median Contig | Assembly Likelihood ($-\log_{10}$) |
|--|-----------------------|-----------------|------------------------|---------------|--------------------------------------|
| Q_{30} and Bulk | 127,508 | 83,264,944 | 851 | 411 | $-2.832 \cdot 10^{10}$ |
| Q_{30} and Bulk with Khmer processing | 147,902 | 92,395,841 | 789 | 423 | $-1.224 \cdot 10^9$ |

Table 4.4.7: Trinity assemblies (with $-\min\text{-kmer-cov} 2$) of Q_{30} trimmed, taxonomically selected single cell and bulk libraries with and without Khmer digital normalisation and K-mer abundance filtering. Khmer pre-processing improved the assembly likelihood by an order of magnitude, and significantly increased the total size of the assembly while only having a marginally negative effect on contig N₅₀s.

As Khmer pre-processing both significantly improved assembly run time as well as the overall assembly quality (as assessed in the Trinity assembly comparison metrics above table 4.4.7) digitally normalised and K-mer abundance filtered bulk and taxonomically selected Q_{30} trimmed SCT were determined to be the optimal pre-processing for this dataset.

4.4.3 ASSEMBLY

4.4.3.1 REFERENCED

Referenced assembly using the divergent *Chlorella NC64A*, *Coccomyxa subellipsoidea C-169*, *Tetrahymena thermophila*, *Paramecium caudatum* genomes as references was largely ineffectual. Of all bulk and SCT reads only 0.3 and 0.4% mapped to the algal references respectively. Similarly, only 0.6 and 0.9% of reads mapped to the related ciliate genomes. This level of mapping is on the order of random chance. Of the read which mapped, a high proportion (73 – 82%) mapped non-uniquely. This suggests mapping was occurring in low complexity regions and is a statistical artefact for the most part instead of biological significance.

The addition of gene junction annotation files for the reference genomes to improve spliced mapping only improved the percentage of reads mapping by 0.05 – 0.3 percentage points. With so few reads mapping, any attempt to class transcripts from this using cufflinks resulted in 10 – 23 total transcripts.

Therefore, referenced assembly using divergent related genomes proved impossible for this dataset.

4.4.3.2 DE NOVO ASSEMBLY

The results of the initial assembler comparison using the Q₃₀ trimmed taxonomically selected SCT libraries (Light1-9, Light1-10, Light1-11, Dark1-2, Dark2-3, Dark2-6, Dark2-8) and bulk libraries are shown in table 4.4.8.

| Assembler | Parameters | Number of Contigs | Bases Assembled | Assembly Likelihood – log |
|------------------|------------------|-------------------|-------------------------|---------------------------|
| SOAPdenovo-Trans | K ₂₃ | 374,325 | 7.64 · 10 ⁷ | 3.778 · 10 ¹⁰ |
| | *K ₆₄ | - | - | - |
| | *K ₈₀ | - | - | - |
| TransAbyss | K ₂₀ | 3,272,137 | 1.722 · 10 ⁸ | - |
| | K ₃₂ | 853,079 | 1.321 · 10 ⁸ | - |
| | K ₆₄ | 376,280 | 9.755 · 10 ⁷ | - |
| | Merged | 3,055,851 | 2.71 · 10 ⁸ | -3.113 · 10 ¹⁰ |
| Oases* | - | - | - | - |
| IDBA-tran | - | 54,113 | 2.7 · 10 ⁷ | -4.589 · 10 ¹⁰ |
| IDBA-MTP/UD/MT** | - | - | - | - |
| Trinity | min_kmer_cov 2 | 127,508 | 8.326 · 10 ⁷ | -2.832 · 10 ¹⁰ |
| Bridger | K ₂₅ | 114,582 | 9.707 · 10 ⁷ | -2.587 · 10 ¹⁰ |

Table 4.4.8: *De novo* assemblies of Q₃₀ trimmed taxonomically selected single cell libraries and bulk libraries (but not digitally normalised or K-mer abundance filtered) with a range of assemblers and parameters. K-mer size used for assemblers with that option are indicated in the Parameters column e.g. K₂₃ indicates a 23-mers. Bridger and Trinity outperformed other assemblers in terms of assembly likelihood and rational contig numbers and sizes. * indicates assemblies programs that failed to run to completion due to insufficient computational resources (despite using a server with 500GB of memory) ** indicates assemblies which failed due to coding errors in the application.

Critically, Oases, the IDA-MTP/UD/MT pipeline and SOAPdenovo-Trans at higher K-mer values all failed to run to completion correctly with the dataset. In the case of Oases and SOAPdenovo-Trans at higher K-mer values this was due to exhaustion of system memory and in the case of IDBA-MTP/UD/MT workflow an unresolved coding error resulting in repeated segmentation faults.

However, Trinity and Bridger both consistently generated assemblies of approximately equal size (100-130,000

contigs of rational sizes: N50s of 700-850 and mean and median contig sizes of 600-660 and 410-470) across a variety of assembly parameters (not shown). Furthermore, they both consistently generated the assemblies with the greatest likelihoods (from RSEM-Eval), and ran most computationally efficiently.

Trinity and Bridger assemblies using digitally normalised and K-mer abundance filtered, taxonomically selected, Q₃₀ trimmed, single cell and bulk libraries performed even better in terms of assembly likelihood and read incorporation.

| Assembler | Parameters | Contigs | Bases Assembled | Assembly Likelihood (– log) |
|------------------|-----------------------------|----------------|------------------------|------------------------------------|
| Bridger | K ₁₉ | 102,686 | $8,209 * 10^7$ | $-1.729 * 10^9$ |
| | K ₂₅ | 113,106 | $9.866 * 10^7$ | $-1.183 * 10^9$ |
| | K ₃₁ | 112,391 | $8.941 * 10^7$ | $-1.143 * 10^9$ |
| Trinity | Minimum K-mer Coverage of 1 | 176,097 | $1.113 * 10^8$ | $-1.214 * 10^9$ |
| | Minimum K-mer Coverage of 2 | 147,902 | $9.239 * 10^7$ | $-1.238 * 10^9$ |

Table 4.4.9: Assembly summaries of Q₃₀ trimmed taxonomically selected SCT and bulk reads after digital normalisation and K-mer abundance filtering. Parameters used in the assembly indicates any special parameter settings used in the assembly i.e. K₁₉ indicates a K-mer size of 19 was used.

Smaller K-mer values (19-mer) performed worse in the case of the Bridger assembly with the optimal assembly in terms of contig number and size was the K-mer size of 25. This was slightly lower in terms of likelihood than the 31-mer Bridger assembly. The digitally normalised and filtered Trinity assemblies generated much larger assemblies overall but still produced good likelihoods.

4.4.3.3 ASSEMBLY COMBINATION

Two assemblies were combined using the tr2aacds.pl script in EvidentialGene and a minimum CDS size of 100. The first consisted of all successfully completed assemblies of non-normalised/filtered reads i.e. SOAPdenovo-Trans, TransAbyss (multiple K-mer assembly merged using built-in tool), IDBA-tran, Trinity and Bridger in table 4.4.8. The second, of the 3 Bridger digitally normalised assemblies and two Trinity assemblies described in table 4.4.9.

| Assembly | Input Contigs | Collapsed Contigs | Assembly Likelihood (– log) |
|-----------------------------------|----------------------|--------------------------|------------------------------------|
| Non-normalised Assemblies | 3,726,379 | 46,063 | $-4.347 * 10^{10}$ |
| Normalised Assemblies | 652,182 | 53,628 | $-1.823 * 10^9$ |
| CD-HIT 90% meta-clustering | 99,691 | 94,628 | $-5.133 * 10^{10}$ |

Table 4.4.10: Summary of merged multi-assemblies. Collapsed contigs is the number of contigs found in the merged set by the EvidentialGene pipeline. The level of assembly reduction and redundancy removal is high and, at first appearance, is impressively consistent between meta-assemblies despite differences in preprocessing. However, CD-HIT metaclustering at 90% identity shown at the bottom demonstrated that there was very little overlap between these two minimised assemblies. Even the merged normalised assemblies generated a meta-assembly of lower overall likelihood than the best individual constituent assemblies.

The combination of all non-normalised assemblies produced a surprisingly small set of contigs, however, also both assemblies also had lower likelihoods than any of their constituent assemblies. It is of interest that despite both generating similar numbers of contigs there was next to no overlap between the two combinations as assessed by clustering using CD-HIT at a similarity of 90%.

Therefore, the assembly selected for downstream binning and analysis was Bridger assembly of bulk and library screened single cell normalised and K-mer abundance filtered reads with a K-mer size of 31 as it displayed the best likelihood while maintaining assembly statistics within expected ranges.

4.4.4 BINNING

4.4.4.1 ORF CALLING

From the 112,391 contigs in the final selected assembly (31-mer Bridger Normalised and Taxonomically Selected SCT and Bulk) - 1,005,370 ORFs longer than 30 amino acids were identified using a “Tetrahymena” encoding. Using the 500 longest of these ORFs to train a Markov Model and removing shorter ORFs that lay entirely within a longer ORF resulted in a final set of 70,605 ORFs.

4.4.4.2 PERFORMANCE OF BLAST-BASED BINNING

10,000 of these ORFs were randomly selected and used to search the NCBI nr database with BLASTP with an expectation of $1e - 5$. Based on the taxonomic provenance of the top-hit these ORFs were assigned to a particular originating bin. The initial identification and binning of recovered transcripts into host and endosymbiont categories was tested using this phylogenetic approach. The results of this analysis is plotted below. This demonstrates that the initial bin identifications were accurate for endosymbiont ($\sim 92\%$) and food ($\sim 94\%$) derived transcripts.

4.4.4.3 PHYLOGENY-BASED BIN CLASSIFICATION

The 70,605 transdecoder called peptide sequences were then run through the automatic phylogeny generation pipeline (“Dendrogenous”) against the 40 representative genomes described above. Of these, 38,193 had no BLAST hit against any genome database sequence and thus were not used to generate phylogenies. A further 9,335 had less than 4 hits and thus were not used to generate phylogenies but were taxonomically sorted based on the BLAST hit binning criteria to give 8,574 “host” sequences, 258 “endosymbiont”, 395 “food” and 108 “unknown”. An additional 9 sequences had insufficient numbers of sites when masking to generate a phylogeny (≤ 30). Finally, 10 phylogenies were malformed due to a latent bug in FastTree2. Therefore, 22,672 phylogenies were successfully generated and named from the input sequences.⁴

The training dataset and test datasets were visualised to ensure that the training dataset (generated during a previous iteration of these analyses) was representative of the test dataset. These plots demonstrate a possible under representation of “Unknown” and/or “Food” samples (fig. 4.4.10) but do reflect a training dataset that largely encompasses a good quantity of the same feature space as the test dataset (fig. 4.4.11).

⁴However, in speed testing “Dendrogenous” did prove very efficient at rapidly generating phylogenies with its fully parallelised mode capable of generating 100 phylogenies randomly selected transcripts against 41 genomes in an average 2:22.50 minutes. The same pipeline run serially took an average of 23:41.39 minutes and the stage-wise parallel was very marginally faster at an average of 21:45.02 minutes.

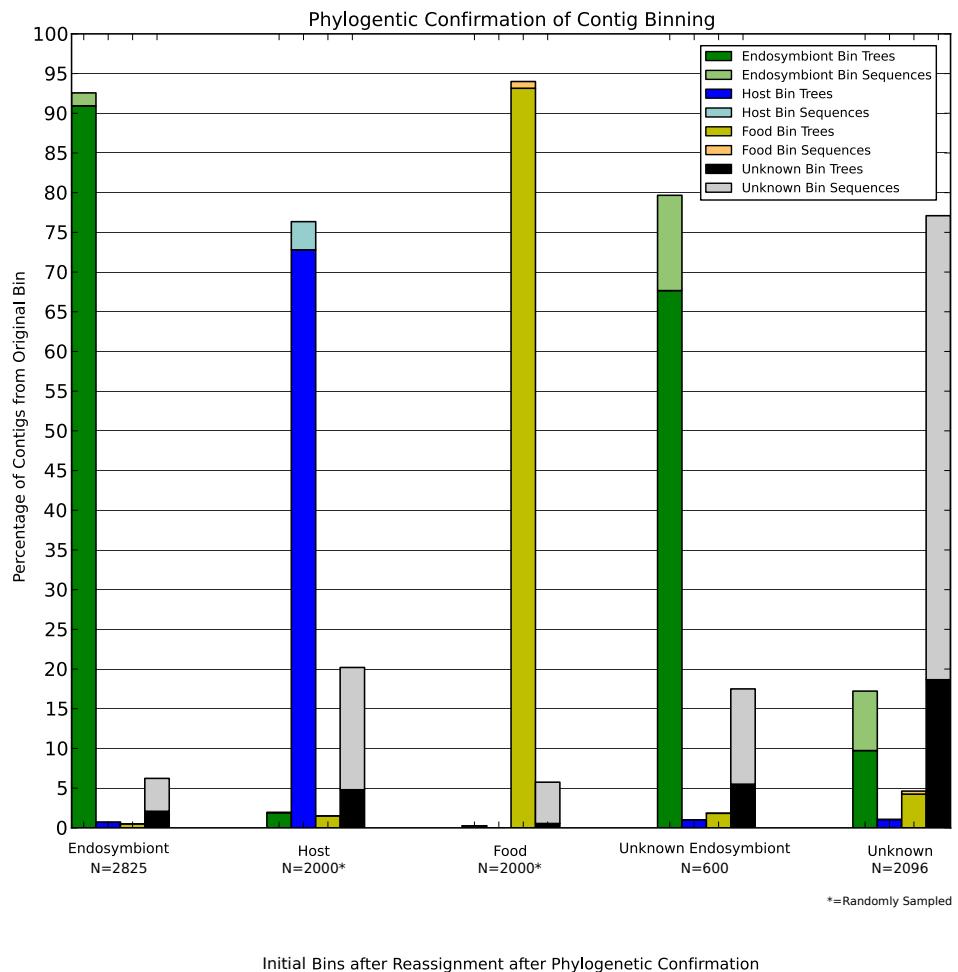


Figure 4.4.9: Preliminary analysis of change in binning after manual phylogenetic confirmation. This analysis was based on an earlier iteration of the assembly and ORF calling.

The large range of classification algorithms were fitted to this training dataset and hyperparameters were efficiently optimised using random search and Bayesian optimisation on the cross-validation folds. The average F-1 scores across classes were tallied and compared revealing K-Neighbours the most effective classification algorithm for this dataset (fig. 4.4.12).

As can be seen in the confusion matrix (and manual parsing of the classification reports from each classifier (see appendix <++>)) K-neighbours (like the majority of classifiers) poorly classified “Unknown” samples but largely performed well (0.89 – 0.9 for each class (table 4.4.11)).

When the trained K-Neighbours model was used to classify the unlabelled 22,672 phylogenies: 415 were “endosymbiont”, 2253 “unknown”, 19476 “host” and 531 “food”.

Therefore, of the 70,095 called ORFs in total there were: 28,050 were “host” derived, 673 “endosymbiont”, 40446 “unknown” and 926 “food”.

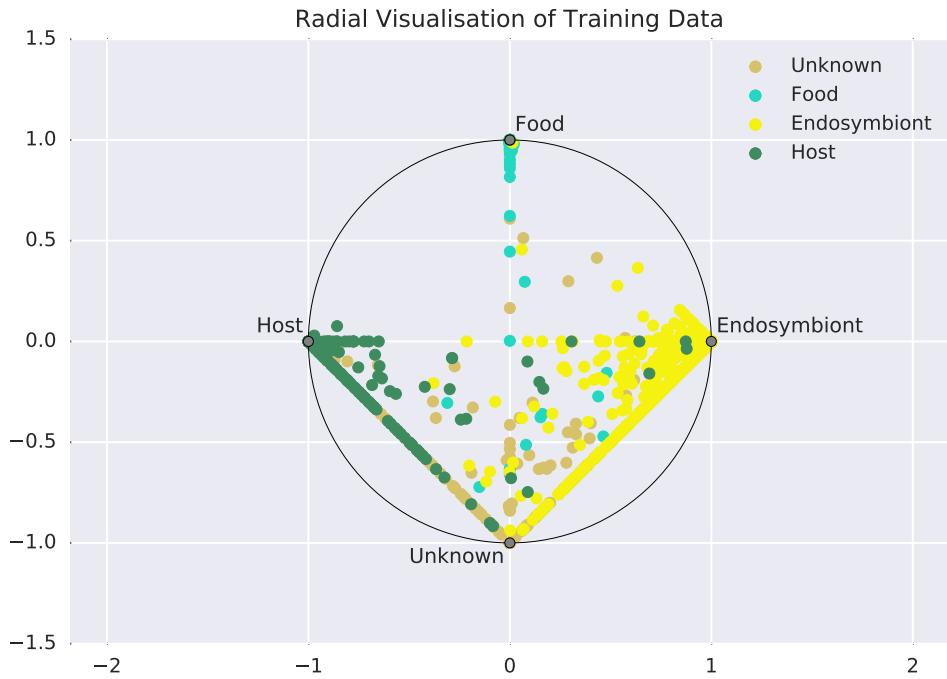


Figure 4.4.10: Radial Visualisation of Manually Parsed Training Data. All input features are normalised to unit magnitudes. Each point represents a single training sample (i.e. phylogeny) and its relative proximity to the cardinal points of the unit circle represents the number of closely related taxa considered part of that “class”. Unknown and Food classes can be seen to be particularly problematic and poorly partitioned. represents the

| Label | Precision | Recall | F1-Score | Support |
|-----------------|-----------|--------|----------|---------|
| “Unknown” | 0.96 | 0.84 | 0.90 | 156 |
| “Food” | 0.98 | 0.99 | 0.99 | 426 |
| “Host” | 0.90 | 0.99 | 0.98 | 787 |
| “Endosymbiont” | 0.97 | 0.99 | 0.89 | 359 |
| average / total | 0.95 | 0.96 | 0.95 | 1728 |

Table 4.4.11: Classification report of a trained and optimised K-Neighbours Classifier using a leaf size of 30, minkowski distance metric and 50 neighbours. Note the poor performance on “Unknown” samples but generally good ($\geq 90\%$) on other labels. This can likely be explained by the “miscellaneous” nature of this label and the diverse phylogenies that comprise it.

4.4.4.4 PERFORMANCE RELATIVE TO TAXASSIGN

TAXAssign performed relative poorly at taxonomic classification/binning of transcripts. Of 70,605 CDS sequences only 2,043 (2.893%) were assigned a phylum level taxonomic identity (table 4.4.12).

Of these top level assignments:

This can be contrasted with the 29649/70605 or 41.99% classified using the phylogeny and supervised classification system..

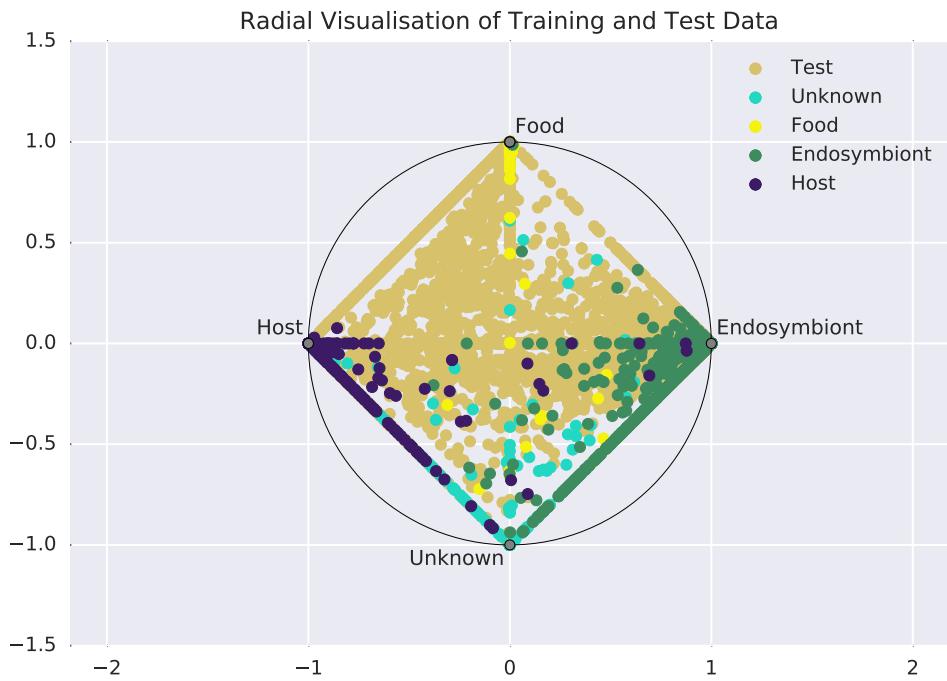


Figure 4.4.11: Radial Visualisation of Test Data and Training Data. All input features are normalised to unit magnitudes. Each point represents a single training sample (i.e. phylogeny) and its relative proximity to the cardinal points of the unit circle represents the number of closely related taxa considered part of that “class”. Test shows the position of all unlabelled phylogenies. This plot shows where the training data is poorly sampled – specifically phylogenies that only contain “host” and “food” taxa or “host” and “endosymbiont” taxa. These phylogenies may prove problematic to easily classify.

4.5 DISCUSSION

4.5.1 LIBRARY SCREENING IS A KEY STAGE IN sc-RNASEQ

Despite evidence and hope that nanoscale methods can greatly reduce levels of contamination (Blainey and Quake, 2011), the taxonomic profiling conducted here indicates a high level of bacterial (and viral) contamination in the scRNA-Seq. Therefore, much as library contamination is one of the key issues with single cell genomics (Blainey, 2013; Lusk, 2014), it is also highly important in SCT. This is in concordance with the findings of (Kolisko et al., 2014), in which enigmatic, bacterial contamination was a problem in single cell eukaryotic transcriptomes.

Single cell methods are particularly prone to contamination issues from reagents, laboratory environment and enigmatic nucleic acids within the biological samples themselves. This is due to the low-input concentration and high amplification necessary in these approaches (Blainey, 2013) leading to enrichment of non-target sequences, especially bacterial contaminants present around or within the *P. bursaria* host. It is critical to identify and discard highly contaminated libraries in *de novo* assemblies as contaminant reads severely complicate the assembly graph thus increasing the computational difficulty and reducing the accuracy of the de-Bruijn graph path resolution. This was highlighted by observations in the preliminary stages of this project that the inclusion of certain (SCT) libraries would increase assembly run-time and lead to the generation of fragmented transcripts relative to assemblies without those libraries.

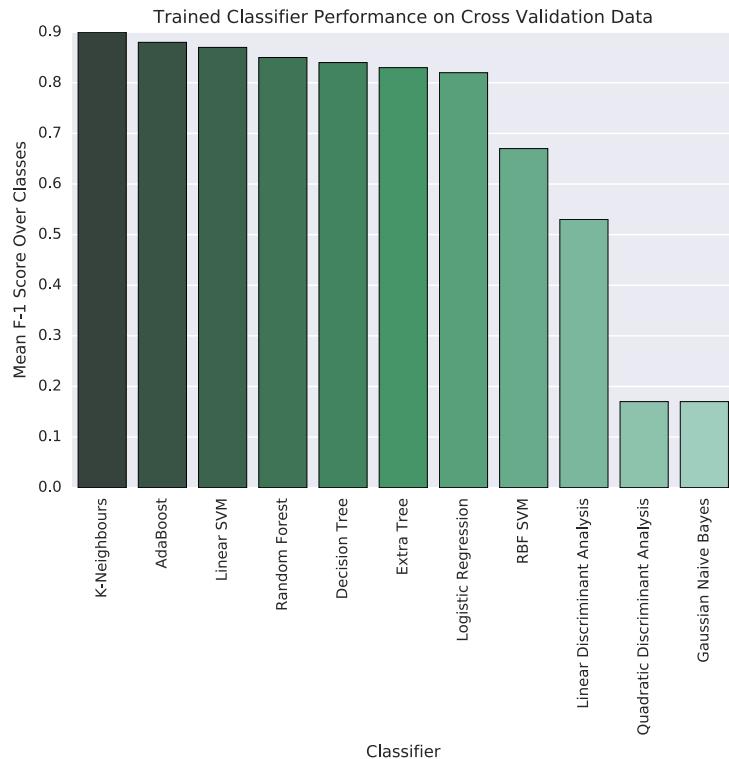


Figure 4.4.12: Classification report must be straightforward - a report of P/R/F-Measure for each element in your test data. In Multiclass problems, it is not a good idea to read Precision/Recall and F-Measure over the whole data any imbalance would make you feel you've reached better results. That's where such reports help.

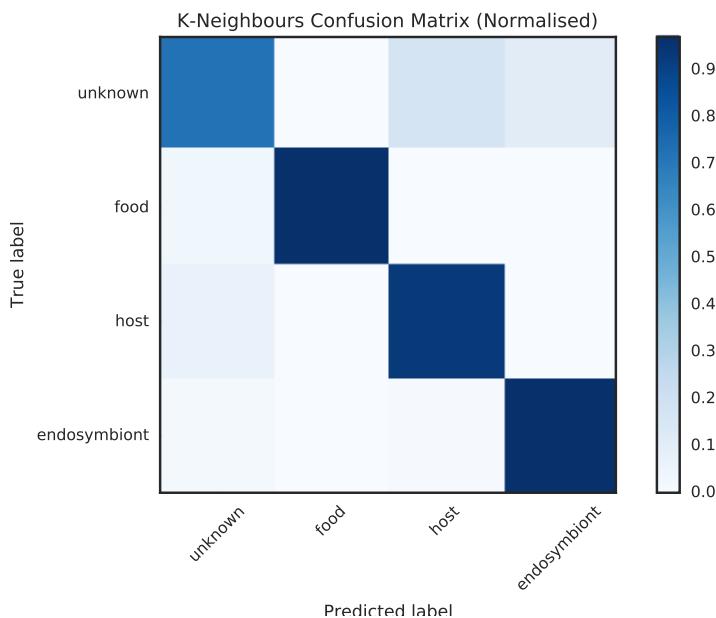


Figure 4.4.13: Normalised Confusion Matrix for K-Neighbours. These plots highlight the problematic classes in the cross-validation dataset. The heatmap corresponds to the proportion of samples classified with a given predicted label compared with their true labels. "Host" samples are accurately classified however a small number are erroneously classified as "Unknown". Similarly, "Unknown" samples are relatively poorly classified in general.

Taxonomic profiling also reveals other features of the dataset that wouldn't necessarily be obvious otherwise. For instance, our profiles revealed systematically low levels of Viridiplantae related reads across both bulk

| Class | Phylum | Sequences Assigned |
|--------------------------|------------------------|--------------------|
| “Host” “Endosymbiont” | Intramacronucleata | 97 |
| | Streptophyta | 101 |
| | Chlorophyta | 58 |
| | Cyanobacteria | 1 |
| | Proteobacteria | 1270 |
| | Firmicutes | 80 |
| | Actinobacteria | 35 |
| | Bacteriodetes/Chlorobi | 29 |
| | Chordata | 365 |
| “Food” | Chlorovirus | 94 |
| | Arthropoda | 7 |
| “Unknown” | | |

Table 4.4.12: Phylum level TAXAssign assignments for (2,043/70,605) CDS called from the Bridger 31-mer assembly. Only 2.893% were assigned using this method relative to 39.72% for the phylogenetic supervised learning (Dendrogenous-Arboretum) method. Therefore, this demonstrates the how well this method works relative to conventional binning approaches like TAXAssign.

and sc-RNAseq libraries and in both lit and dark conditions. Despite care being taken to ensure thorough lysis of the chitinous *Micractinium* cells during RNA extraction and a ratio of $\sim 300 : 1$ endosymbiont to host cell ratio this may be related to lytic inefficiencies (Korfhage et al., 2015) or potentially just relatively fewer endosymbiont transcriptomic activity relative to host and associated bacteria species. Finally, it is possible that due to the endosymbiont being largely provisioned by the host it may be relatively transcriptionally inactive and thus relatively fewer transcripts can be recovered.

Intriguingly, taxonomic profiling of the bulk libraries showed a very low percentage of reads mapping to any sequence in the nr protein database. This was significantly lower than the sc-RNAseq libraries. While this finding is concerning it is likely to be an artefact of the older sequencing platform the bulk data was generated on. These paired reads were sequenced via the GAI and were half the length of the HiSeq2500 reads used for the SCTs. Shorter reads and a relatively higher technical error rate on this platform may have played a role in this marked decline in recognisable reads. Despite this the relative proportion of bacterial reads to eukaryote reads among the recognisable reads is much lower for bulk libraries than SCT. This does however suggest, that read length is highly important to accurate contamination screening/taxonomic profiling.

Taxonomic profiling of reads/libraries proved surprisingly robust to trimming. The profiles generated in “DueyDrop” were largely identical regardless of whether the input library had been trimmed or not (even at high stringency thresholds). This indicates that screening can occur before the computationally expensive trimming process without loss of accuracy. While the results presented here demonstrate that the profile generated from a relatively small random subsample of a library is largely consistent with that of the profile generated from the library as a whole.

“DueyDrop” could potentially be improved in such a way that an entire library can be screened quickly instead of just subsamples. Briefly, this would involve quantifying exact K-mer matches between library reads and a pre-generated database of known taxa using efficient probabilistic hashing data structures such as a bloom filter, or more likely count-min sketch and an efficient k-mer counting library such as Jellyfish (Marçais and Kingsford, 2011). This would have a major speed advantage compared to the BLASTX-based method of “DueyDrop” and would thus allow entire libraries to be checked in reasonable time. A similar approach to this has been implemented

as a service by <https://www.onecodex.com/>, although this focuses specifically on screening for medically relevant taxa. However, it would require laborious workarounds to handle K-mers shared between multiple taxa in the database, translation of reads and/or database sequences into a matching sense and form (e.g. protein), and use of a locality-sensitive hash function to handle scenarios where there is no exact k-mer match. This latter issue is particularly problematic for libraries consisting of transcripts from poorly sampled sections of the tree of life where exact matches would become commensurately rarer as sampling sparsity increases. While still affected by this problem, the BLASTX/Diamond approach implemented in the “DueyDrop” scripts are relatively more robust to these problems due the explicit probabilistic modelling of sequence divergence built into the BLAST alignment algorithm (e.g. E-values). Other potential improvements to “DueyDrop” would be to incorporate some degree of automatic clustering of phylogenetic profiles using unsupervised learning and possibly manifold embedding, potentially including a form of anomaly detection to discover contaminated libraries. Robustness of taxonomic inference for each profile can also be improved by taking all hits instead of just the top one and resolving conflicting phylogenetic signal using a lowest common ancestor algorithm over all the hits.

4.5.2 COMBINING SINGLE CELL AND BULK TRANSCRIPTOME DATA CREATES NEW CHALLENGES

Contrary to previous studies in optimising conventional RNA-seq assemblies where: permissive trims (Macmanes, 2014), error correction (Macmanes and Eisen, 2013; Macmanes, 2015) and the combination of multiple assemblies (Nakasugi et al., 2014) have been demonstrated as effective tactics in recapitulating a comprehensive set of transcripts *de novo*, MDA-based SCT and bulk datasets such as the one investigated above exhibit different properties. It is important to pick a trimming threshold which minimises sequence error (mostly substitutions (Yang et al., 2013b)) as these result in assembly of spurious sequences but doesn't discard reads necessary to complete transcripts (Macmanes and Eisen, 2013; MacManes, 2014). For the *P. bursaria* - *M. reisseri* bulk and SCT dataset the optimal trimming threshold - based on both mapping statistics to preliminary assemblies and final assembly likelihoods proved to be a harsh threshold of Q₃₀.

Following a similar theme, despite numerous indications that error correction is important for improving the accuracy of genomic and transcriptomic assemblies using Illumina reads (e.g. (Molnar and Ilie, 2014; Macmanes, 2015)) in the case of this dataset error correction appeared to be have a minimal effect with few reads being corrected and downstream assemblies being largely equivalent to those without error correction. In fact, permissively trimmed (Q₅), error corrected assemblies were shown to have lower likelihoods and smaller assembly sizes than more conservatively trimmed assemblies (Q₃₀). It should be noted that SEECER, while an RNAseq specific error correction tool is not optimised for single cell MDA-based datasets, and Bayeshammer is not optimised for transcriptomic data. Therefore, the poor performance of error correction in this dataset might be a consequence of the lack of error correction tools designed for MDA-based sc-RNAseq datasets. It will likely prove beneficial, as datasets of this type become more prevalent, to develop tools for this specific use case combining the most effective features of the MDA aware BayesHammer and the RNA-seq optimised SEECER. This said, there are several other available Illumina RNA-Seq error correction tools which were not trialled, “SEECER” was chosen in accordance to the recommendations based on dataset and hardware heuristics (> 50M reads and the availability

of a high memory system) ([Macmanes, 2015](#)) but as we've demonstrated the limitations of such heuristics on new types of data it might be worth investigating these alternative tools further.

Finally, merging multiple assemblies proved a sub-optimal strategy with all merged assemblies generating lower likelihood assemblies than the best individual assembly. While not merging assemblies might mean specific transcripts might not have been recovered, especially assemblies at a range of K-mer sizes (short K-mers generally recover lower expression transcripts and vice versa for long K-mers), the much higher likelihoods meant the best performing individual assembly (Bridger using 31-mers, Q₃₀ trimmed taxonomically selected SCT and bulk libraries) was preferred.

These results suggest that MDA-based single cell transcriptomic datasets do not behave in a qualitatively similar way to bulk RNA-seq in terms of pre-processing and assembly parameters. This means care must be taken incorporating advice and heuristics derived from studies based on analysis bulk RNA-seq datasets (e.g. ([Macmanes and Eisen, 2013](#); [MacManes, 2014](#); [Macmanes, 2015](#); [Nakasugi et al., 2014](#))). As further studies using MDA SCTs are completed (e.g. ([Kolisko et al., 2014](#))) a greater understanding of the optimal analysis of this type of dataset will emerge.

4.5.3 PRE-ASSEMBLY READ PARTITIONING IS NON-TRIVIAL

As we expect the PbMr metatranscriptome to contain predominantly a highly AT-rich organism, *Paramecium*, (ranging from 24.1 to 28.2%GC in *P. aurelia* species complex and *P. caudatum* ([Aury et al., 2006](#); [McGrath et al., 2014](#))) and a very GC-rich organism, *Micractinium*, (*Chlorella variabilis* NC64A genome is approximately 67.1%GC, the highest found in a sequenced eukaryote genome (in 2010) ([Blanc et al., 2010](#))), the utility of pre-assembly read partitioning was assessed. This GC pattern was supported by the clear bimodal GC distribution that can be observed in fig. 4.4.1. However, under careful observation it was apparent that the bimodal GC distribution was more attributable to the presence of a GC-rich contaminant such as Rhizobiales ([Peralta et al., 2011](#)). Therefore, in practice pre-read partitioning was mainly attempted to try to remove these contaminant reads from screened libraries. Theoretically, accurate pre-assembly read partitioning could transform a complex assembly graph into two relatively simpler assembly tasks. As well as simplifying path resolution accuracy, if this method could speed up assembly considerably and thus allow more iterations to optimise other assembly parameters.

This pre-assembly partitioning has been tried with mixed success in other meta-omic analyses (e.g. ([Dröge and Mchardy, 2012](#))). However, a lack of fast efficient tools to accomplish this led to the creation of “parKour”. The developed GC partitioning package proved very effective at rapidly and relatively computationally efficient clustering of PE RNAseq data. ParKour generated clusters with centroids reasonably where they may be expected from inspection of per read GC probability densities (see fig. 4.4.3) i.e. partitioning out the GC rich potentially contaminant reads likely from *Rhizobia* bacterial species. Unfortunately, in the case of this dataset clustering proved ineffective at improving assembly accuracy and fully removing groups of contaminant reads with large GC skews. The likely explanation for this is that even 150bp paired end reads are too short to consistently statistically demonstrate the GC-AT bias of the originating organism. This means any partitioning is likely to remove a significant number of reads necessary to complete transcripts due to local variation in AT bias. The high number of shorter

contigs is indicative of the kind of assembly fragmentation that would be expected in this situation.

However, the relative efficiency and theoretical benefits of this type of clustering indicates there may be some potential to utilising a similar but less naive approach in future work. It may be possible to combine “DueyDrop” and “ParKour” to all read-level screening and partitioning of reads on the basis of taxonomic profile and compositional characteristics such as GC% and tetramer frequencies. This could improve resolution of clusters and decrease the observed contig fragmentation effect while performing accurate taxonomic screening.

Other improvements could include the consideration of alternative clustering algorithms such as k-medoids ([Kaufman and Rousseeuw, 1987](#)) with more robust outlier stability or large scale database clustering algorithms such as DBSCAN ([Ester et al., 1996](#)) or BIRCH ([Zhang et al., 1996](#)) (allowing non-convex clusters). Silhouette coefficients and analysis⁵ ([Rousseeuw, 1987](#)) can be incorporated to aid determination of the expected number of clusters when it cannot be determined *a priori* from inspecting the data as well as validation of generated clusters. Unfortunately, other validation and analysis systems are somewhat limited due to the lack of ground truth labelling available. Alternatively, a variational Bayes approach could be implemented to determine the optimal number of clusters (e.g. CONCOCT ([Alneberg et al., 2014](#))). Finally, memory efficiency can be improved by use of streaming clustering algorithms (e.g. those discussed in ([O’Callaghan et al., 2002](#))) in which all data does not require to be loaded into a matrix at a given time and be clustered as they are parsed.

4.5.4 DIGITAL NORMALISATION GREATLY IMPROVES ASSEMBLIES

Digital normalisation, a method to remove redundant read data from libraries and thus reduce the computational burden of assembly ([Brown et al., 2012](#)), was also investigated for this dataset and found to be a highly effective strategy in improving assemblies of mixed bulk and MDA SCT data.

Interestingly, some have argued that error correction is special case of general digital normalisation ([Krasileva et al., 2013](#)). This is supported by the fact that many error correction algorithms operate on similar principal of attempting to remove low abundance K-mers from input datasets. K-mers with a low abundance are more likely to be due to sequencing errors than representing novel biological diversity. This said digital normalisation has the potential to spuriously discard true variation that is merely undersampled in our libraries due to the high level of contamination.

This hypothesis is somewhat supported by the disproportionate retention of bulk reads relative to noisy single cell reads. However, within the context of the single cell reads the more highly contaminated Dark reads were retained in roughly equal proportion to the light reads. Suggesting, MDA derived data may also just display a greater quantity of low-level sequencing error.

It should be noted that the digitally normalised and K-mer abundance filtered assemblies also incorporated more bases overall than the equivalent assembly using the full libraries therefore resultant assemblies were just high confidence (and likelihood) subsets of the initial input data.

One factor that has not been adequately analysed in the context of this work is that of sequencing depth. Future

⁵ $s_s = \frac{b-a}{\max(a,b)}$ where s is Silhouette coefficient, a is the mean distance between a sample and all other points in the same class and b is the mean distance between a sample and all other points in the next nearest cluster ([Pedregosa et al., 2011](#)). Therefore, s is a measure of cluster definition.

studies will need to carefully consider sufficient sequencing depth given the noise and prevalence of contamination in MDA based data.

4.5.5 ASSEMBLY AND ASSEMBLY ASSESSMENT

While we have demonstrated that some progress can be made identifying optimal pre-processing parameters using measures such as mapping metrics it is very difficult to identify the parameters (preprocessing or otherwise) which will lead to the “best” *de novo* assembly without actually generating the assembly. Assembly can be considered an example of Wolpert and Macready’s “No Free Lunch Theorems” (Wolpert and Macready, 1995, 1997) as (in the case of *de novo* assembly) it is fundamentally a hamiltonian/eulerian cycle search problem (equivalent in the de-Bruijn formulation) and therefore any two assembly implementations (in different assemblers and/or with different parameters) should ultimately be equivalent across all possible input datasets.⁶. For this reason, it is necessary to try assembly using a random of assembly parameters and indeed a range of both *de novo* and referenced assemblers.

Unfortunately, the task of identifying the “best” *de novo* transcriptome assembly is also a non-trivial task (Neil and Emrich, 2013). Many widely used assembly assessment metrics have been shown to be inconsistent measures in simulated sequencing data, especially those metrics related to individual contigs (theoretically different transcript splices). Metrics such as average length and N₅₀ prove consistent across both simulated sequencing depth and read lengths i.e. they improve towards (Neil and Emrich, 2013). Furthermore, the number of possible metrics is greatly reduced if assessment is mainly conducted in a reference-free manner (Li et al., 2014). As the majority of assemblies were *de novo* and the suitability of the related but divergent genomes proved lacking it was necessary to restrict to reference-free assembly assessments. Therefore, a model-based reference-free assembly scoring algorithm (RSEM-EVAL (Li et al., 2014) was, along with standard (if imperfect) metrics, used to evaluate different assemblies in this study. The assumption of the accuracy of the RSEM-EVAL likelihood is a strong one, and deserves careful re-consideration in further work.

In terms of referenced assembly using divergent relatives, it is safe to conclude that despite other findings that even divergent (up to 15%) genomes can generate transcriptomes of higher-quality than *de novo* (Vijay et al., 2013), the potential references are too divergent in the case of the PbMr to be of any utility.

Overall, a comparison of *de novo* assemblies using a range of assemblers and parameters on “optimally” pre-processed read data demonstrated the clear superiority of both “Bridger” and “Trinity” Trinity comes with the advantages of being a generally better developed tool that interlocks effectively via several plugins and utility scripts with other tools and analysis pipelines.

However, despite being relatively newer and consisting of a less mature and tested codebase, Bridger proved to be a slightly more effective assembly tool overall. Unfortunately, coding problems and a lack of public active development means it is non-trivial to successful use this tool. In the process of implementing the above analyses it was necessary to fix several bugs present in the assembler. These upgrades were merged into the code and are

⁶This should be taken with a pinch of salt, a proof of this theorem applied to the case of assembly is beyond both the scope of this thesis and my abilities

available on GitHub https://github.com/fmaguire/Bridger_Assembler. Hopefully, by rehosting this code on a public development and collaboration platform (as well as adding continuous integration) will spur further development of this promising tool.

Interestingly, despite strong evidence supporting the need to combine assemblies due to the size of the disjoint sets of transcripts recoverable from different algorithms and parameter choices (Lowe et al., 2014) assembly merging systematically led to worse overall assemblies with this dataset (as assessed by RSEM-EVAL likelihood scores). The likelihood of the merged assemblies were worse than the best individual constituent assemblies.

4.5.6 BINNING

However, even once a good assembly has been generated it is still necessary to identify the likely originating species of a given transcript i.e. host, endosymbiont, food bacterial contaminant or other contaminant. While a successful partitioned pre-assembly strategy may simplify this process it would still be sensible to confirm bins using downstream analyses that use full length assembled transcripts and thus more potential data than are present in shorter individual PE reads. Rough, approximate bins were generated using a simple "top BLAST hit" approach following ORF calling (using Tetrahymena and Universal encodings) against a set of representative predicted proteomes. In order to assess how accurate these bins were likely to be, 10,000 were randomly selected and rapid maximum-likelihood phylogenies were generated using the transcript sequence as a seed to sample the entire RefSeq protein nr database. This was accomplished using "Dendrogenous", a rewritten and modified version of a pipeline originally known as "Darren's Orchard" which first appeared in (Richards et al., 2009). Phylogenies were manually assessed to check whether the resultant topology was congruent with the BLAST based binning i.e. are supposedly "endosymbiont" transcripts branching principally with archaeplastida taxa. However, due to the slow largely manual nature of this phylogeny assessment process it would be infeasible to repeat this for all transcripts generated from a single assembly, let alone investigating several.

Therefore, this became a fundamental classification problem with the 10,000 manually verified phylogenies forming a handy training dataset for supervised learning. To determine the best performing classification algorithm and hyperparameters for this dataset an automated search was conducted using bayesian optimisation. This was then converted to a binning script named "Arboretum". High throughput phylogeny generation, parsing and supervised classification is a more sensitive and powerful way in which to bin transcripts into their likely originating organisms provided a reasonable level of *a priori* knowledge of the system at hand. This demonstrably operates better than established although simpler approaches such as TAXAssign or top BLAST hit. While, classification accuracy (and F-1) is sub-optimal for "Food" and "Unknown" bins it (table 4.4.11) a decent level of precision and recall for the target bins of "Host" and "Endosymbiont".

However, this classification is still a work in progress and could potentially be improved by the addition of anomaly detection in place of the catch-all (and subsequently poorly classified) "Unknown" classes. Furthermore, there are several potential possible improvements in the classification itself that could be made. Specifically, unsupervised clustering pre-training could potentially forgo the need for laborious manual generation of the training dataset and minimise the difficulties in handling these aberrant phylogenies. An AutoML bagging estimator such

| Transcript Bin | Number of Transcripts | Called ORFs |
|-----------------------|------------------------------|--------------------|
| Endosymbiont | 8,975 | 4,275 |
| Host | 18,793 | 17,920 |
| Food | 18,516 | - |
| Unknown | 66,107 | - |

Table 4.6.1: Summary of transcriptome assembly and binned sequences

as one of those implemented in the AutoML project ([Eggensperger et al., 2013](#)), a variational autoencoder pre-processing following by a deep neural network classifier or using a phylogeny specific kernel function ([Vert, 2002](#)) in a Gaussian process or SVM system all offer potential algorithmic improvements. Finally, incorporation of additional sequence related features such as K-mer coverage and composition into each samples may help greatly improve the fidelity of classification.

4.6 CONCLUSION

In conclusion, for this dataset the optimal pre-processing was determined to be careful taxonomic screening of input libraries, followed by trimming at a high (Q₃₀) threshold and subsequent digital normalisation and low-abundance K-mer filtering. The optimal assemblies were generated using larger (25-31 K-mer) sizes and utilised the Bridger (and to lesser extent Trinity) assembly algorithms. While pre-assembly read partitioning proved ineffective in this implementation, in future a less naive method that incorporated both read-level taxonomic data and compositional information could potentially improve assemblies of complex eukaryotic metatranscriptomes, especially those that combined bulk and single cell RNAseq data. Generally, MDA-based single cell datasets are noisy and difficult to work with. Potentially, they require the existence of robust references to aid assembly or a much greater depth of sequencing.

Finally, I have demonstrated that BLAST based transcript binning alone is ineffective at accurately binning transcripts. Fast, automated phylogeny generation and the subsequent use of supervised learning (particularly large ensemble models such as Random Forests and those AutoML algorithms) can potentially improve the quality of such binning. Further work in the implementation of unsupervised clustering of generated phylogenies could conceivably forgo the laborious process of manually generating a training dataset.

Points arising in this analysis:

- It is possible to generate a functional working transcriptome combining bulk and MDA based RNAseq (see table 4.6.1)
- sc-RNAseq libraries generated from dark samples are problematic.
- Binning methodologies may prevent easy finding of novel genes due to a lack of homology.

"All models are wrong, but some are useful"

- George E.P. Box & Draper: *Empirical model-building and response surfaces*, 1987

5

Metabolic integration

5.1 INTRODUCTION

The linking of metabolism between host and endosymbiont is a fundamental stage in endosymbiotic integration (Bhattacharya et al., 2007; Karkar et al., 2015). Complementation of the respective metabolic deficiencies/limitations in host and endosymbiont allow exploitation of novel niches and provide the key selective benefits of endosymbiosis (Hoffmeister and Martin, 2003).

In order to identify putative metabolic integration it is necessary to identify the primary “points of contact” between the metabolic networks of host and endosymbiont. These “points of contact” comprise two major classes of proteins, transporters and secreted proteins. Specifically in *P. bursaria* system, host and endosymbiont transporters which localise to the perialgal vacuole (PV) membrane and the outer-membrane of the endosymbiont, and proteins which are secreted by host and endosymbiont into PV lumen.

Another way to investigate metabolic integration is the annotation and analysis of known metabolic pathways in host and endosymbiont from binned transcriptome sequences. This allows identification of pathways being expressed while in the endosymbiotic relationship and potentially indicates further likely sites of integration.

Finally, a characterisation of metabolites present in the system can be used to further analyse host and endosymbiont metabolic function. A metabolomics based-approach has the benefit of directly characterising the metabolites themselves and thus interrogating biological activity at its level of operation. Therefore, relative to

analyses relying on abstracted measures such as transcription levels or gene copy number, novel information about cellular dynamics can be revealed. This is particularly important in cases of of cryptic regulatory systems that break-down direct mapping from genes to transcripts to protein to metabolites. By using both targeted and untargeted chromatography and mass spectrometry metabolomics approaches it is possible to survey the combined pool of host and endosymbiont both qualitatively and quantitatively. These inferences can then be correlated with other data such as the presence of transcripts involved in the synthesis or degradation of these metabolites.

By utilising 3 separate streams of metabolic analysis: comparative transcript annotation and mapping, directed identification and analysis of transporters and secreted proteins and metabolomics, we maximise the strength of any inferences and reduce the chance of false negatives preventing the identification of biologically important metabolic integration.

First, I will summarise what is currently known about the metabolic integration of *P. bursaria* and its green algal endosymbionts, before briefly discussing the nature and identification of evidence of further metabolic integration.

5.1.1 METABOLISM OF HOST AND ENDOSYMBIONT

The most obvious point of metabolic integration in any endosymbiosis featuring a photosynthetic partner is that of the flow of photosynthates from endosymbiont to host.

This is believed to primarily be in the form of carbohydrates such as maltose (Muscatine, 1967) In return, the host facilitates increased rates of photosynthesis in the endosymbiont (Sommaruga and Sonntag, 2009), via supply of CO₂ (Parker, 1926), one or several forms of nitrogen (Johnson, 2011), and mono- and divalent cations such as K⁺, Mg²⁺, and Ca²⁺. All of which have key roles in photosynthesis (Kato and Imamura, 2009b).

The transfer of maltose, glucose, fructose and malate from endosymbiont to host has been observed previously using radiolabelling e.g. (Brown and Nielsen, 1974). Furthermore, green algae strains competent to form endosymbioses can be induced via modifying pH to release significantly more photosynthate (in the form of ~ 95% maltose) than strictly free-living strains (on the order of 5.4 – 86.7% vs. 0.4 – 7.6% of total photosynthate) (Muscatine et al., 1967). Interestingly, export of photosynthate from the PV to the host cytoplasm may be dependent on a transporter derived from the *C. variabilis* 1N in its respective *P. bursaria* endosymbioses (Kodama and Fujishima, 2008).

In terms of the uptake of saccharides *C. variabilis* F36-ZK endosymbiont strains seem incapable of directly utilising sucrose, maltose, glucose or fructose in free-living culture (Kamako et al., 2005; Kato and Imamura, 2009b). Glucose, does promote growth however, via an apparent sensing pathway that leads to the up-regulation of amino acid uptake (Kato and Imamura, 2009b). On the other hand, the free-living *C. vulgaris* strain has an inducible system for active hexose uptake (Tanner et al., 1974). In order of highest to lowest uptake rate the free-living *C. vulgaris* took up sucrose, glucose and maltose but not fructose (Kato and Imamura, 2009b).

Nitrogen is the most transferred material between host and endosymbiont after carbon (Kato and Imamura, 2009b). There has been considerable research and interest in exactly what form this nitrogen exchange takes (Kato et al., 2006; Kamako et al., 2005; McAuley, 1986).

C. variabilis (both NC64A and the Japanese F36-ZK) have been found to be able to use amino acids effectively as a nitrogen source but only minimally utilise ammonium (NH_4^+) and incapable of effective nitrate (NO_3^-) or nitrite (NO_2^-) use (Kamako et al., 2005; Kato and Imamura, 2009b). Similar patterns have been observed in *M. reissleri*, although all strains tested could utilise nitrate and 3/4 could use nitrite to greater or lesser degrees (Kessler and Huss, 1990). On the other hand, free-living species such as *Parachlorella kessleri* can effectively utilise all of the nitrogen sources mentioned (Kato and Imamura, 2009b) (although amino acid utilisation has to be induced with glucose treatment (Cho et al., 1981)).

In terms of amino acids as a nitrogen source, there unfortunately, isn't a high degree of correlation between the ability to uptake an amino acid and its usage. Even rate of uptake does not indicate utilisation (Kato and Imamura, 2009b). For example, while *C. variabilis* F36-ZK can uptake all 20 amino acids, only 6 (L-arginine, L-asparagine, L-glutamine, L-serine, L-alanine) were found to promote growth (Kato et al., 2006). This was despite some of these 6 being taken up at lower rates than some non-utilised amino acids such as L-proline, L-cysteine or L-leucine (Kato et al., 2006).

Similarly, *C. variabilis* NC64A was found to have stimulated growth in the presence of L-arginine and L-glutamine, whereas another *C. variabilis* strain, 3N813A, used every amino acid apart from L-lysine and L-glutamic acid (McAuley, 1986; Kato and Imamura, 2009b).

The free-living *C. vulgaris* NIES-227 on the other hand was found to not utilise any amino acid apart from low levels of uptake of L-arginine (Kato et al., 2006). Therefore, even within the *C. variabilis* strains there is a range of traits in terms of amino acid uptake and utilisation.

Kinetic analyses and competitive assays indicate 3 amino acid transport systems in *C. variabilis* F36-ZK, a general amino acid transporter for all amino acids except L-alanine, a basic transporter for L-arginine and L-lysine and a specialised L-alanine transporter (Kato and Imamura, 2009a,b). All of these are constitutively functioning, active, amino-acid symporters (Kato and Imamura, 2009a,b).

As *P. bursaria* cannot import nitrate (Albers et al., 1982) the heterogeneous loss of nitrate and nitrite utilisation in several endosymbiotic strains is perhaps not surprising (Kato and Imamura, 2009b). As there is no pressure to maintain enzymes necessary for this pathway when an endosymbiont this may explain the appearance. This reduced selection pressure for nitrate and nitrite utilisation when in the endosymbiont milieu may explain the presence of low-activity mutant Nitrate Reductase (NR) and Nitrite Reductase (NiR) in the two *C. variabilis* strains (Kato and Imamura, 2009b).

The final major group of host-endosymbiont transferred metabolites are those of mono- and divalent cations. Specifically Ca^{2+} , Mg^{2+} and K^+ (Kato and Imamura, 2009b). All of these have key roles in photosynthesis. Interestingly, Ca^{2+} has also been found to inhibit amino acid uptake (Kato and Imamura, 2008). As glucose has been found to increase amino acid uptake, some researchers have hypothesised that the relative concentration of photosynthates and Ca^{2+} in the PV lumen defines a photosynthate-amino acid barter system between endosymbiont and host (Kato and Imamura, 2009b).

5.1.2 IDENTIFYING DIRECT POINTS OF CONTACT

5.1.2.1 TRANSPORTER PROTEINS

The most important group of proteins in the control and evolution of metabolic integration is that of host and endosymbiont transporter proteins. This is due to their role in facilitating diffusion and active transport across the lipid membranes that exist between host and endosymbiont.

Without transporters many metabolically important large uncharged polar molecules (e.g. carbohydrates, amino acids) and charged molecules (e.g. the various biologically relevant cations and anions such as H^+) are incapable of significant rates of diffusion across membranes even in the presence of high concentration gradients. Therefore, the presence of transporters is vital to facilitating any interaction involving these groups of metabolites.

Several metabolites are, however, capable of unfacilitated diffusion across lipid membranes at significant rates. These include important respiratory gases such as O_2 and CO_2 , hydrophobic compounds like benzene and small uncharged polar molecules (e.g. H_2O and ethanol) (Cooper, 2013; Alberts, 2015). Despite this, transporter proteins have evolved to facilitate even more rapid diffusion of some of these metabolites e.g. aquaporins (Agre et al., 1993).

Finally, certain transporters can provide the ability to actively transport metabolites against concentration gradients. This involves the expenditure of energy (typically in the form of ATP) to directly pump compounds or generate an opposing gradient which can be exploited (primary vs secondary active transport).

There are 5 functional groups of transporters (Saier et al., 2014):

- Channel/Pore types which catalyse diffusion of metabolites along concentration gradients e.g. porins and the Mitochondria and Plastid Porin (MPP) family.
- Electrochemical Potential-driven transporters which use a carrier-mediated process to catalyse uniportation (single metabolite) or cotransportation (two species in the same direction, symportation, or two species in opposite direction, antiportation). These make use of chemiosmotic gradients but generally do not directly make use of cellular energy molecules such as ATP. However, many make use of a gradient/potential generated by the active transport of solutes by another complex, in this case they can referred to as secondary active transporters. Electrochemical Potential-driven transporters are a very large family and include the ubiquitous major facilitator superfamily (MFS) and Cation Diffusion Facilitator (CDF) families.
- Primary active transporters which use a direct source of chemical, electrical or light energy such as ATP, voltage or photon to drive transport against concentration gradients. Transporters of this type form many of the components fundamental to life as they allow an organism to decouple itself directly from environmental and intracellular gradients. They include rhodopsins, ATP-binding Cassette (ABC) Superfamily, and the general Secretory Pathway (Sec) family.
- Group translocators, which modify a substrate during transportation e.g. polysaccharide synthesis during

secretion in the Polysaccharide Synthase/Exporter family and the Fatty Acid Group Translocation (FAT) family which can acylate fatty acids during transport.

- Transmembrane Electron Carriers which transport single electrons from a donor to an acceptor across a membrane. The major groups of these include the cytochrome and Photosystem I complexes.

This analysis will focus on transporters of the first 4 classes. In the case of *P. bursaria* and its endosymbiont we are particularly interested in the host perialgal vacuole membrane and the outer membrane of the green algal endosymbiont in the case of *P. bursaria*.

Therefore, the first step to the successful analysis of the metabolic integration of host and endosymbiont is accurate identification of transporter proteins present in their respective binned transcriptome sequences. By identifying both the identity of these proteins and qualitatively investigating their day:night expression targets can be generated for further analysis, validation, and proof of localisation to the PV and algal outer membrane.

Transporter proteins can be identified primarily via annotation and homology searches to previously identified transporters (Saier et al., 2006, 2009, 2014) and direct identification of transmembrane (TM) domains motifs. All transporters feature at least 1 TM helix, usually considerably more (von Heijne, 2006). However, as not necessarily every sub-unit of a transporter will contain a TM domain and the presence of partial transcripts in our assemblies it is necessary to not rely totally on TM prediction to discover transporter proteins.

One obvious target of such is the sugar and amino acid transporters which facilitate the secretion and uptake of photosynthate and amino acid between host and endosymbiont.

5.1.2.2 SECRETED PROTEINS

The next major class of proteins involved in endosymbiosis are those which are secreted in the PV lumen. Host and endosymbiont derived proteins targeted to the lumen of this vacuole are responsible for generating the local environment for endosymbiosis. This is fundamental to the exchange of metabolites between host and endosymbiont as this is the cellular context in which that exchange must occur. Furthermore, in the establishment of endosymbiosis secreted effectors from the endosymbiont are likely responsible for the modification of a phagosomal vesicle into the perialgal vacuole.

One particularly interesting example to identify would be the hypothesised endosymbiont derived transporter exported to the PV membrane that is responsible for export of photosynthate to the host cytoplasm (Kodama and Fujishima, 2008).

The identification of secreted proteins relies on the analysis of signal peptides and/or a range of standard classification algorithms based on sequence and compositional features. Signal peptides are short 15-30 amino acid N-terminal sequences that determine targeting of proteins to cellular compartments (Schatz and Dobberstein, 1996; Rusch and Kendall, 1995).

SignalP (Nielsen et al., 1997) has proven the most effective method of predicting the presence of signal peptides (Lee et al., 2009; Petersen et al., 2011). This method uses a standard feed-forward artificial neural network with 8-41 hidden units (depending on whether the organism is eukaryotic, gram positive or gram negative) trained

with back-propagation.

Subcellular localisation of a given protein can also be inferred using tools such as the WoLF PSORT (Horton et al., 2007). This tool implements a standard K-neighbours classifier trained on localisation labelled proteins from SwissProt and uses PSORTII (Nakai and Kanehisa, 1992; Nakai and Horton, 1999; Horton and Nakai, 1997) and iPSORT (Bannai et al., 2002) derived sequence features and automatic inference of weightings (Horton et al., 2006).

In addition to these tools, there are many other systems that have been designed to predict protein localisation. In order to maximise the probability of successful identification of secreted proteins I have created a consensus ensemble classifier that incorporates predictions from several of these tools, principally SignalP, TMHMM, TargetP, ChloroP, and WoLFPSort.

5.1.3 METABOLIC MAPPING

Metabolic pathways form the functional backbone of all biological processes. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999; Okuda et al., 2008; Kanehisa et al., 2014) and MetaCyc (Caspi et al., 2007) databases form an important resource for contextualising genomic and transcriptomic scale results into these networks. The utility of this approach is emphasised by the presence of numerous tools and analytical pipelines to explore these databases e.g. (Okuda et al., 2008; Nakao et al., 1999; Karp et al., 2002, 2010; Antonov et al., 2008; Klukas and Schreiber, 2007)

By comparing the relatively complete predicted peptides sets from the sequenced endosymbiotic algae *C. variabilis* NC64A and *Coccomyxa subellipsoidea* C-169 genome projects to the predicted endosymbiont binned peptides from the transcriptomes of *M. reisseri* and *C. variabilis* 1N it is possible to infer what the endosymbiotic metabolic pathways are active in the latter two endosymbiont. Furthermore, identification of pathways unique to individual algae can identify potential distinct adaptations to endosymbiosis in that algae.

5.2 AIMS

The principal aim of this chapter is to identify likely points of metabolic integration between host and endosymbiont to generate targets for subsequent targeted mass spectrometry, RNAi and qPCR based validation experiments.

This will be achieved by:

- Identifying transporter proteins present in the endosymbiont binned transcripts from the CCAP1660/12 RNA-Seq analysis and analysing them for qualitative expression across day and night.
- Identifying secreted proteins present in the endosymbiont binned transcripts from the CCAP1660/12 RNA-Seq analysis and analysing them for qualitative expression over day and night.
- Comparative analysis of metabolic pathways between host and endosymbiont relative to sequenced green algal genomes.

- A pilot untargeted global metabolomic profile of the system and comparison of day to night.
- Targeted quantitative analysis of amino acid concentrations between day and night.

5.3 METHODS

5.3.1 TRANSPORTER ANALYSIS

5.3.1.1 TRANSPORTER IDENTIFICATION PIPELINE

Transporters were identified in the 4 sets of sequences (*C. variabilis*, *M. reisseri*, *C. vulgaris* and *C. subellipsoidea*) using the following set of pipe-lined filters:

1. Transmembrane (TM) domains were predicted for each sequence using an HMM approach implemented as part of TMHMM2 (Sonhammer et al., 1998; Krogh et al., 2001) and sequences predicted to contain at least 1 TM domain were extracted.
2. These sequences were then used to search a PFAM database of profile HMMs (Eddy, 1995) via HMMER3's hmmscan utility (Eddy, 1995; Johnson et al., 2010; Eddy and R, 2011; Mistry et al., 2013) and sequences with a hit to a PFAM domain at an independent E-value of $1e^{-5}$ were retained.
3. These hits were then finally filtered for PFAM domains which mapped to transporter families classified by the Transporter Classification Database (TCDB) (Saier et al., 2006, 2009, 2014) mapping files.

Additionally, to ensure thorough discovery of all *M. reisseri* transporters *M. reisseri* binned sequences were BLASTP-ed against the nr protein database with an e-value of $1e^{-3}$ and a maximum of 20 hits. InterproScan (Zdobnov and Apweiler, 2001) was then used to further annotate these proteins incorporating results from BlasProDom (Servant et al., 2002), FPrintScan (Attwood et al., 1994), HMMER (Mistry et al., 2013) scans against the PIR (Barker et al., 1998), PFAM (Bateman, 2002), SMART (Schultz et al., 1998), PANTHER (Thomas, 2003) and TIGRFAM databases (Haft, 2003), ProfileScan (Gribskov et al., 1988), HAMAP (Lima et al., 2009), PatterScan, SuperFamily (Gough and Chothia, 2002), SignalP (Petersen et al., 2011), TMHMM (Sonhammer et al., 1998), Gene3D (Buchan et al., 2002), Phobius (Käll et al., 2007) and Coils. Results were then mapped to GO terms (Ashburner et al., 2000; Harris et al., 2004) and annotated via BLAST2GO (Conesa et al., 2005).

Finally, all proteins annotated to have a GO term associated with “transport” and “transport activity” specifically GO:0006810 and its child terms were extracted.

This set was then filtered for the presence of TM domains and to remove organelle related transporters such as cytochrome system and photosystems as these are unlikely to localise anywhere other than the mitochondria and chloroplast.

5.3.2 SECRETOME PREDICTION

A conservative set of secreted proteins were predicted using the following consensus ensemble classifier:

1. Signal peptides were detected using SignalP4.1 and mature sequences created for each sequence with a signal peptide
2. Sequences detected to have a TM domain (by TMHMM) within either the mature sequence or full length for proteins without signal peptides were discarded.
3. Signal peptides were re-added to mature sequences and the remaining sequences were then filtered using for those predicted as secreted by TargetP1.1
4. These sequences were then filtered down to those which had extracellular localisation in WoLFPSORT 0.2
5. Finally, for the endosymbiont, secreted protein found to have a Chloroplast targeting signal (via ChloroP1.1) were removed.

A permissive set was also generated without the TMHMM filtering and for all proteins predicted to be extracellular or have a signal peptide targeting for secretion.

5.3.3 QUALITATIVE EXPRESSION ANALYSIS

Kallisto ([Bray et al., 2015](#)) was used to pseudoalign and estimate abundances for all taxonomically screened single cell libraries (4 dark and 3 light) to the called “endosymbiont” binned CDS sequences from the *P. bursaria* CCAP 1660/12 transcriptome (see Chapter 2).

Kallisto doesn’t align reads to references in the same manner as conventional short read alignment algorithms such as Bowtie2 ([Langmead and Salzberg, 2012](#)). Instead of specifically mapping a read to a set of co-ordinates it instead determines which transcripts are compatible with the alignment of a given read. This is achieved via decomposition of transcripts in de-Bruijn graphs and fast k-mer hashing to compare reads to transcript graph nodes in constant time. These k-compatibility classes are then used with bootstrapped expectation-maximisation to estimate transcript quantification and determine uncertainty ([Bray et al., 2015](#)).

Results were visualised and analysed using “sleuth” and the seaborn plotting library ([Waskom et al., 2015](#)).

5.3.4 METABOLIC MAPPING ANALYSIS

First, predicted proteomes were obtained or generated for *Coccomyxa subellipsoidea* C-169, *Chlorella variabilis* NC64A, *Chlorella variabilis* 1N, and *M. reisseri*.

For *M. reisseri* the endosymbiont binned sequences from the transcriptomic sequencing project discussed in the previous chapter were used. *Coccomyxa subellipsoidea* C-169 genome project ([Blanc et al., 2012](#)) version 2.0 JGI annotated proteins (created 12-01-2014) were downloaded from JGI’s Phytozome v10.3.1 ([Goodstein et al., 2012](#)). Similarly, the “best” annotated proteins from version 1 of the *Chlorella variabilis* NC64A genome project ([Blanc et al., 2010](#)) were downloaded from JGI’s genome portal ([Grigoriev et al., 2011; Nordberg et al., 2014](#)).

However, to obtain *C. variabilis* 1N endosymbiont peptides a reassembly and binning of raw sequencing data from ([Kodama et al., 2014](#)) was conducted (discussed below).

Once all sequences were acquired they were annotated using KEGG orthology. This was achieved using the KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007) single-directional best hit with both BLAST and GHOSTZ (Suzuki et al., 2014, 2015) method against the following 40 gene sets: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Glycine max*, *Vitis vinifera*, *Oryza sativa*, *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Micromonas* sp., RCC299, *Cyanidioschyzon merolae*, *Galdieria sulphuraria*, *Saccharomyces cerevisiae*, *Candida albicans*, *Neurospora crassa*, *Aspergillus nidulans*, *Coccidioides immitis*, *Schizosaccharomyces pombe*, *Ustilago maydis*, *Encephalitozoon cuniculi*, *Monosiga brevicollis*, *Dictyostelium discoideum*, *Acanthamoeba castellanii*, *Plasmodium falciparum* 3D7, *Cryptosporidium hominis*, *Tetrahymena thermophila*, *Paramecium tetraurelia*, *Phaeodactylum tricornutum*, *Emiliania huxleyi*, and *Guillardia theta*.

KEGG annotations were then plotted onto KEGG metabolic networks and compared to identify key aspects of differences between the algal species and host and endosymbiont metabolic.

5.3.4.1 CHLORELLA VARIABILIS 1N ASSEMBLY

232.3M 100bp paired-end reads from (Kodama et al., 2014)'s bulk RNAseq transcriptome of *Paramecium bursaria* Yad1g1N (syngen 3, mating type 1) bearing *Chlorella variabilis* 1N endosymbionts were downloaded from the DNA Data Bank of Japan (DDBJ) (Tateno et al., 2002; Kaminuma et al., 2011) in Sequence Read Archive (SRA) format (Leinonen et al., 2011; Kodama et al., 2012) (accession DRA000907 (Kodama et al., 2014)).

These reads were then converted to fastq using “fastq-dump” using the SRA Toolkit (National Center for Biotechnology Information, 2011). Reads were then trimmed for sequencing adapters using ILLUMINACLIP and SLIDINGWINDOW with a window size of 4 and a minimum average quality of 5 in Trimmomatic (Bolger et al., 2014).

Reads were then error-corrected using “SEECER” with a k-mer size of 25 and default settings otherwise (entropy of 0.6 and a cluster log-likelihood of -1) (Le et al., 2013). Error-corrected reads were digitally normalised using a K-mer size of 25 and a coverage of 20 (Brown et al., 2012) and low abundance K-mers in high coverage reads were filtered (Zhang et al., 2014, 2015) using the Khmer software package (Döring et al., 2008; Crusoe et al., 2015).

Assemblies were completed in a modified/fixed version of Bridger 2014-12-01 (Chang et al., 2015) (available at https://github.com/fmaguire/Bridger_Assembler) and Trinity v2.0.6 (Grabherr et al., 2011; Haas et al., 2013) both with K-mer sizes of 25.

An alternative Trinity assembly was also completed using SLIDINGWINDOW Q30 trimmed reads without normalisation or error correction.

Assemblies were compared using RSEM-EVAL (Li et al., 2014) and the best overall assembly selected on the basis of likelihood. ORFs were called from the best assembly using universal and tetrahymena encodings via TransDecoder (Haas et al., 2013) retaining the best scoring sequences and those with HMMR hits to PFAM and BLASTP hits to the SwissProt database.

Phylogenies were generated for each sequence using the same approach and pipeline described in Chapter 4. These phylogenies were subsequently classified using the same trained K-Neighbours supervised learning al-

gorithm. Any sequence that didn't have enough BLAST hits in the genomes used to generate a phylogeny (*s*) were parsed based on what hits were retrieved. Those with no hits were classified as "unknown" and those with hits were classified based on the origin of those hits e.g. hits to green algae and plant genomes were considered "endosymbiont" and so on.

Finally, the ORF bins for host and endosymbiont from both encodings were manually combined and reconciled to generate transcript bins. With the transcripts binned into "host" and "endosymbiont" ORFs were recalled from them using the appropriate encodings.

5.3.5 METABOLOMICS

3 mass spectrometry analyses were conducted to investigate the presence/absence and relative abundances of polar and non-polar metabolites. Specifically, GC-QTOF to principally identify metabolites such as carbohydrates, LC-QTOF in positive and negative ionisations to profile general metabolites.

Finally, a targeted mass spectrometry analysis was conducted using LC-QQQ to quantitatively assess the concentration of free amino acids in the host-endosymbiont system during the day and night.

5.3.5.1 UNTARGETED LC-QTOF PROFILING

Sample preparation for mass spectrometry followed standard protocols. Briefly, *s* biological replicates were sampled from *P. bursaria* CCAP 1660/12 cultures at the midpoint of both the day and night cycles. Samples were then dried, flash-frozen in liquid nitrogen, and homogenised using a cell disruptor.

For each sample, 10mg was dissolved in 400 μ l of a solution of 80% MeOH containing 7.2mgml⁻¹ of an umbelliferone internal standard. This solution was kept on ice and vortexed for 30s every 10 minutes for 30 minutes. Samples were sonicated in ice cold water for 15 minutes and then centrifuged at 13k rpm for 10 minutes. Retaining the supernatant in a separate tube, the pellet was resuspended in 400 μ l 80% MeOH and vortexing, sonication and centrifugation steps repeated. The two supernatants were combined and filtered through a 0.2 μ m syringe filter (Chromacol). Samples were sub-divided into two a 5 μ l of each was loaded into an Agilent 1200 Series HPLC with a 3.5 μ m, 2.1 × 150mm Eclipse Plus C18 Agilent column. One sample was then analysed using a positive electron spray ionisation and the other a negative ionisation on an Agilent 6520 accurate mass quadrupole time of flight (QTOF) mass spectrometer. Data was captured using the standard Agilent data acquisition software and converted from ".d" format to the open mzXML format (Pedrioli et al., 2004). The same process was repeated for 7 blank samples under both ionisation conditions as well as an apigenin QC standard.

Samples were analysed primarily using the XCMS R package (Smith et al., 2006; Tautenhahn et al., 2012b). Peak features were detected in each samples using the centWave algorithm with a 30ppm tolerated m/z deviation, minimum peak width of 10 and maximum peak width of 60 (Tautenhahn et al., 2008). Peaks were aligned with a 0.025 m/z width overlap, and a maximum bandwidth 5 retention time deviation. Using the aligned peaks, the retention time deviation between samples were calculated. The samples were then realigned correcting for retention time deviation and integrated using fillPeaks.

In order to determine differential presence of globally detected metabolites unpaired Welch's *t*-tests were conducted comparing the 5 day samples to the 5 night samples. Welch's tests were used as they don't assume equal sample sizes or variances between the two groups (Welch, 1947)¹

P-values from this were corrected for multiple comparisons using false discovery rates (FDR). FDR is a less conservative correction than, the classic family-wise error rate correction, Bonferroni adjustment² but allow maintenance of a greater proportion of statistical power with a slightly elevated risk of Type-I errors.

Features were then annotated using the METLIN metabolite database (Smith et al., 2005a; Sana et al., 2008; Tautenhahn et al., 2012a). Extracted ion base peak chromatograms were manually inspected for each significantly expressed feature and any that weren't clear distinct peaks were discarded. Any sample without an annotation against METLIN was similarly discarded. Finally, annotations were manually parsed and samples with impossible annotations (e.g. chemotherapeutic drugs) discarded.

5.3.5.2 UNTARGETED GC-QTOF PROFILING

Standard sample preparation was conducted with 5 light and 5 dark biological replicates.

Agilent "d" proprietary format outputs were converted to mzXML as above. The majority of the analysis pipeline was conducted using the metaMS R library (Wehrens et al., 2014). Briefly, this involved standard peak picking using the default XCMS algorithm (Smith et al., 2006) followed by clustering into pseudospectra using CAMERA (Kuhl et al., 2012). Pseudospectra were then annotated against the NIST and METLIN databases using a combination of spectral and retention time features. Finally intensity and relative abundances were calculated and tested using FDR t-tests.

5.3.5.3 TARGETED AMINO ACIDS QUANTITATIVE ANALYSIS

5 Day and 5 Night samples were prepared for liquid chromatography using the same approach as the untargeted LC-QTOF analysis. In addition to this, the Day1 and Night1 samples were analysed at both 2x and 0.5x titrations. 3 blank samples were run as well as standards consisting of a complete amino acid mix from Sigma at 0.5 μ M and 4 samples consisting of Asparagine-Glycine-Tryptamine and Leucine-Glutamine-Lysine respectively.

After chromatography samples were ionised using electrospray ionisation and analysed using multiple reaction monitoring optimised for amino acids with an Agilent Technology 6410B enhanced sensitivity triple quadrupole mass spectrometer (QQQ).

Results were analysed using the Agilent Quantitative Analysis software package with peaks normalised in respect to the umbelliferone standard.

¹ $t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + s_2^2 n_2}}$ with degrees of freedom determined via the Welch-Satterwaite-equation: $df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2-1}}$ (Ruxton, 2006)

²P-value threshold α is adjusted relative to the number of comparisons (n). Specifically, significance is determined as a P-value $\leq \frac{\alpha}{n}$.

| Preprocessing | PE Reads |
|--|--------------------|
| Raw Reads | $2.323 \cdot 10^8$ |
| Q₃₀ Trimmed | $1.75 \cdot 10^8$ |
| Q₅ Trimmed | $2.127 \cdot 10^8$ |
| Q₅ Error Corrected | $2.021 \cdot 10^8$ |
| Q₅ Digital Normalisation | $1.09 \cdot 10^7$ |
| Q₅ K-mer abundance filtering | $1.055 \cdot 10^7$ |

Table 5.4.1: Summary of read pre-processing stages for the Kodama library. This table further emphasises the massive amount of redundancy that digital normalisation removes from the assembly. The minimal effect of K-mer abundance filtering is likely due to a relative redundancy in this pre-processing from the error correction step.

| Assembly | Contigs | Likelihood ($-\log$) |
|---|---------|------------------------|
| Trinity Q₅ Normalised | 101,957 | $1.216 \cdot 10^9$ |
| Bridger Q₅ Normalised | 62,504 | $1.285 \cdot 10^9$ |
| Trinity Q₃₀ | 53,938 | $5.619 \cdot 10^9$ |

Table 5.4.2: Summary of resultant assemblies of sequencing data derived from (Kodama and Fujishima, 2014). As the Trinity Q5 Normalised assembly had the best likelihood while generating the greatest number of contigs it was used for downstream analyses.

5.4 RESULTS

5.4.1 *P. BURSARIA-C. VARIABILIS YAD1G1N* ASSEMBLY

A complete *de novo* assembly was conducted of sequencing data from (Kodama and Fujishima, 2014). Based on experiences assembling *P. bursaria*-*M. reisseri* CCAP 1660/12 (see Chapter 4). Briefly, all samples were combined, trimmed to Q₅, error corrected, digitally normalised and abundance filtered. Another dataset was prepared just using adapter trimming and a quality threshold of Q₃₀. Assemblies conducted with these datasets using both Trinity and Bridger *de novo* assemblers were evaluated using RSEM-EVAL and the optimal assembly chosen on score.

Therefore, the optimal assembly chosen for further analysis was the Trinity Q₅ normalised assembly. From the 101,957 transcripts 193,906 ORFs were called using tetrahymena encoding and 20,875 universal. These were subsequently binned using the same approach as used in Chapter 4.

Finally, “Host” and “Endosymbiont” binned transcripts were re-ORF called using the appropriate encodings to result in a host ORF bin of 61,239 sequences and an endosymbiont bin of 5,565 peptides.

This is relatively reasonable dataset sizes especially when considered against the source of the raw sequencing

| Bin | Number of Transcripts |
|--------------|-----------------------|
| Food | 3,873 |
| Endosymbiont | 8,627 |
| Host | 53,295 |
| Unknown | 36,162 |

Table 5.4.3: Summary of transcript binning for the Q5 Trinity assembly of Yad1g1N. A much greater proportion of transcripts were assigned to host and endosymbiont for this assembly than in the single cell based assemblies previously conducted.

| | <i>C. variabilis</i> 1N | <i>M. reisseri</i> CCAP 1660/12 | <i>C. variabilis</i> NC64A | <i>C. subellipsoidea</i> C-169 |
|----------------|-------------------------|---------------------------------|----------------------------|--------------------------------|
| Peptides | 5,565 | 4,275 | 9,791 | 9,629 |
| 1+ TM domains | 695 | 419 | 1,722 | 1,709 |
| 1+ TM and TCDB | 251 | 185 | 690 | 697 |

Table 5.4.4: Summary of the sizes of the complete transporter complements identified in the various algal sequence sets. The two genome based predicted proteomes generated much larger predicted sets of proteins (*C. variabilis* NC64A and *C. subellipsoidea* C-169).

Annotation of *M. reisseri* Binned Peptides

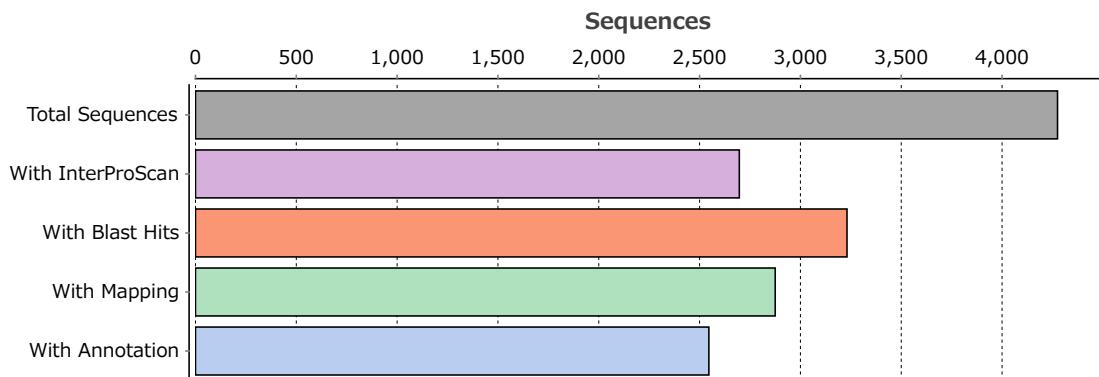


Figure 5.4.1: Summary of Endosymbiont Bin Annotation. A good proportion but not all transcripts were annotated.

data. Specifically, the analysis of (Kodama et al., 2014)), which involved the elimination of endosymbionts from the culture. Therefore, only half of the libraries theoretically contained *C. variabilis* sequences and the recovery of 5,565 peptides is reasonably good.

5.4.2 TRANSPORTER IDENTIFICATION

To identify transporters present in the translated protein dataset of the endosymbiont bins of the CCAP1660/12 and Yad1g1N1 *Paramecium bursaria* transcriptome assemblies, as well as the Chlorella NC64A and Coccomyxa C-169 predicted proteomes the following process was used:

The genome based predictions from *C. variabilis* NC64A and *Coccomyxa* led to a much larger set of candidate transporter proteins than those from the two binned transcriptomes. However, smaller subsets represent all candidate transporters that are actively being transcribed during endosymbiosis rather than just all those present in the genome. Therefore, this is not necessarily problematic and could be symptomatic of the endosymbiont exhibiting a streamlined metabolism inside the host.

The *M. reisseri* predictions were then supplemented using standard annotation pipelines fig. 5.4.1.

From the GO based annotations there were 427 proteins associated with the GO term for transmembrane transport (GO:0006810). These overlapped with approximately half of the TM/TCDB identification (93/185 proteins). 133 of the 427 GO term annotated transporters had at least 1 TM domain (and this 133 contained all 93 shared TM/TCDB and GO based identifications as would be expected).

On manual inspection, a large number of GO based annotations were partial/subunits or could only very

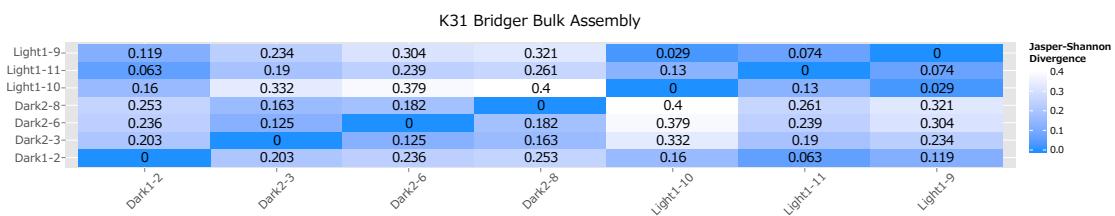


Figure 5.4.2: Comparison of similarity of taxonomically screened single cell libraries. This shows how relatively divergent each library is and demonstrates the high level of noise. The dark and light groups are only very faintly visible. This further supports the use of read-mapping inferences as qualitative rather than quantitative data for MDA-based SCT.

tenuously be referred to as transporters. Due to this only those with a TM domain were retained and added to the list of candidate transporters.

This resulted in a list of 225 putative transporters for *M. reisseri* CCAP 1660/12.

These were then manually filtered to remove obviously organelle related groups, specifically cytochromes/electron transport chain proteins and photosystem related proteins leading to a final set of 161 putative transporter proteins.

In terms of candidate sugar transporters This includes 7 MFS transporters, a hexose sugar transporter, a sugar-phosphate translocator, a sweet1 sugar transporter a nucleotide-sugar transporter, a mannose-6-phosphate isomerase, and a fucose permease. There were also 2 polyol transporters e.g. a glycerol-3-phosphate transporter and an inositol transporter.

Candidate amino acid transporters included 8 amino acid permeases.

5.4.3 SUBSET OF TRANSPORTERS EXPRESSED IN DAY AND NIGHT

Using the nucleotide CDS sequence of the called peptides identified as transporters reads from each of the 7 taxonomically screened single cell libraries were pseudo-mapped to them using Kallisto. Due to the compositional/coverage biases of MDA-based single cell transcriptomics Kallisto statistical inference was likely to be spurious and relate to the well-documented coverage biases of MDA (see fig. 5.4.2). Therefore, a simple presence/absence filter was implemented for the single cell libraries where an estimated Transcripts per million (TPM) was above 0 for at least 1 biological replicate in each condition.

5.4.4 SECRETED PROTEINS

Using the secretome prediction consensus ensemble classifier in conservative and permissive settings resulted in a putative secretome of 24 and 249 proteins respectively. Unfortunately, the permissive group contained a significant number of transporters so the analysis focused on the conservative consensus secreted proteins.

While many of these have no clearly defined function two of the most interesting secreted proteins here relate to carbohydrate metabolism. Specifically, comp65133_seq0|m.57547 which appears to be homologous (expectation of 1.38e-10) to Raffinose synthase, and comp41497_seq0|m.42444 which is a fucosidase.

| Name | Description/Annotation | Top BLASTP Hit Accession | Top Hit Species |
|------------------------|---|--------------------------|--|
| comp1093_seq1 m.1645 | Uric acid-xanthine permease | ref XP_005848091.1 | <i>Chlorella variabilis</i> |
| comp34246_seq0 m.33953 | Uric acid-xanthine permease | ref XP_005848091.1 | <i>Chlorella variabilis</i> |
| comp11781_seq0 m.10145 | Signal recognition particle protein 3 | ref WP_027033724.1 | <i>Chlorella variabilis</i> |
| comp35891_seq0 m.35196 | Sensory protein | ref XP_009389646.1 | <i>Mesorhizobium loti</i> |
| comp13652_seq0 m.12349 | Phagocytic receptor 1b | ref XP_005852067.1 | <i>Musa acuminata</i> |
| comp12191_seq1 m.10601 | Inorganic phosphate transporter | ref WP_028786954.1 | <i>Chlorella variabilis</i> |
| comp23811_seq0 m.24460 | Inorganic pyrophosphatase | ref XP_011399197.1 | <i>Terrimonas ferruginea</i> |
| comp34406_seq1 m.34111 | Potassium transporter 2-like | ref XP_005845784.1 | <i>Auxenochlorella protothecoides</i> |
| comp25848_seq0 m.26580 | Protein trigalactosyldiacylglycerol chloroplastic | ref XP_005842790.1 | <i>Chlorella variabilis</i> |
| comp14064_seq0 m.12814 | Sugar transport protein 10-like | ref XP_005851889.1 | <i>Chlorella variabilis</i> |
| comp15360_seq0 m.14352 | Drug metabolite transporter superfamily | ref XP_005850398.1 | <i>Chlorella variabilis</i> |
| comp16603_seq1 m.16010 | Adenine guanine permease | ref XP_005848680.1 | <i>Chlorella variabilis</i> |
| comp17473_seq0 m.17096 | gprt34 family protein | ref WP_019198042.1 | <i>Afipia birigiae</i> |
| comp18033_seq0 m.17793 | Hypothetical upfo06s protein | ref XP_011398136.1 | <i>Auxenochlorella protothecoides</i> |
| comp21389_seq0 m.21756 | Proton phosphate symporter | ref XP_011396846.1 | <i>Auxenochlorella protothecoides</i> |
| comp22779_seq0 m.23394 | Na^+ solute transporter | ref XP_005846641.1 | <i>Chlorella variabilis</i> |
| comp22990_seq0 m.23026 | Inositol transporter 4-like | gb ESW58454.1 | <i>Pseudomonas fluorescens Bbc6R8</i> |
| comp23135_seq0 m.23792 | MFS transporter | ref WP_026773962.1 | <i>Sediminibacterium sp. OR43</i> |
| comp26434_seq0 m.27099 | ATPase p | ref XP_005844294.1 | <i>Chlorella variabilis</i> |
| comp26454_seq0 m.27109 | Plasma membrane iron permease | ref XP_011400870.1 | <i>Auxenochlorella protothecoides</i> |
| comp27116_seq0 m.2808 | Amino acid permease 6 | ref XP_011400870.1 | <i>Auxenochlorella protothecoides</i> |
| comp27116_seq1 m.2811 | Amino acid permease 2 | ref XP_011400870.1 | <i>Auxenochlorella protothecoides</i> |
| comp27116_seq2 m.2815 | Amino acid permease 2 | ref XP_011400870.1 | <i>Auxenochlorella protothecoides</i> |
| comp30376_seq0 m.30648 | Amino acid permease 2 | gb KPF41572.1 | <i>Rhizobium sp. AAP43</i> |
| comp85197_seq0 m.68822 | Amino acid permease | ref XP_005846503.1 | <i>Chlorella variabilis</i> |
| comp43295_seq0 m.41027 | Lysine transporter? | ref XP_005847284.1 | <i>Chlorella variabilis</i> |
| comp29655_seq0 m.30102 | ABC transporter permease | ref WP_044404984.1 | <i>Rhodopseudomonas palustris</i> |
| comp27137_seq0 m.27822 | ABC transporter ATP-binding protein | gb ACF03145.1 | <i>Rhodopseudomonas palustris TIE-1</i> |
| comp51985_seq0 m.48090 | ABC permease ATP-binding family protein | gb EFD03879.1 | <i>Propionibacterium acnes</i> |
| comp13567_seq0 m.2197 | ABC transporter | ref XP_005849318.1 | <i>Chlorella variabilis</i> |
| comp32752_seq0 m.32669 | Membrane AAA-metalloprotease | ref XP_00584697103.1 | <i>Chlamydomonas reinhardtii</i> |
| comp72377_seq0 m.62224 | ABC transporter permease | ref WP_046827962.1 | <i>Afipia massiliensis</i> |
| comp52708_seq0 m.48773 | ABC transporter permease | ref WP_046827962.1 | <i>Bradyrhizobiaceae bacterium SG-6C</i> |
| comp27304_seq0 m.27991 | Peptide ABC transporter permease | ref XP_005843506.1 | <i>Chlorella variabilis</i> |
| comp83141_seq0 m.53037 | Protein transport protein sec61 subunit alpha isoform partial | ref WP_051503901.1 | <i>Sphingomonas jaspis</i> |
| comp49211_seq0 m.45885 | ATP-binding protein | ref XP_005845268.1 | <i>Chlorella variabilis</i> |
| comp15817_seq0 m.15012 | Transmembrane 9 superfamily member 3 | ref XP_011395771.1 | <i>Auxenochlorella protothecoides</i> |
| comp39178_seq0 m.37743 | Transmembrane 9 superfamily member 3-like | ref XP_005847406.1 | <i>Chlorella variabilis</i> |
| comp49560_seq0 m.46210 | Transmembrane 9 superfamily member 4 | gb AGZ19377.1 | <i>Chlorella sp. ArMoo29B</i> |
| comp29161_seq0 m.29930 | Sulphate transport system | ref XP_005844583.1 | <i>Chlorella variabilis</i> |
| comp35923_seq0 m.35211 | Membrane protein | ref WP_012790048.1 | <i>Chitinophaga pinensis</i> |
| comp36025_seq0 m.35298 | MFS transporter | ref WP_027575203.1 | <i>Bradyrhizobium sp. WSM1743</i> |
| comp39264_seq0 m.37815 | TctA transporter | ref YP_009049981.1 | <i>Chlorella sorokiniana</i> |
| comp40138_seq0 m.38461 | NaDH dehydrogenase subunit 3 | ref XP_005850398.1 | <i>Chlorella variabilis</i> |
| comp40842_seq0 m.39088 | Adenine guanine permease azg1 | ref XP_005843431.1 | <i>Chlorella variabilis</i> |
| comp34747_seq0 m.41450 | Vesicle-associated membrane protein 72d | ref WP_053333544.1 | <i>Gemmimonas phototrophica</i> |
| comp44082_seq0 m.41788 | Serine threonine protein kinase | ref XP_005847889.1 | <i>Chlorella variabilis</i> |
| comp44147_seq0 m.41832 | Calcium-transporting ATPase endoplasmic reticulum-type | ref XP_005848599.1 | <i>Chlorella variabilis</i> |
| comp45947_seq0 m.43343 | Cyclic nucleotide-binding protein | ref XP_005848107.1 | <i>Gemmimonas phototrophica</i> |
| comp46264_seq0 m.43542 | V-type proton ATPase 16 kDa proteoipol subunit | ref XP_005849093.1 | <i>Chlorella variabilis</i> |
| comp46589_seq0 m.43815 | V-type proton ATPase subunit A3-like | ref XP_005845363.1 | <i>Chlorella variabilis</i> |
| comp50679_seq0 m.47029 | ER lumen protein-retaining receptor | ref WP_019199842.1 | <i>Afipia birigiae</i> |
| comp51446_seq0 m.47612 | AMP-dependent synthetase | gb KDD71768.1 | <i>Helicospiridium sp. ATCC 50920</i> |
| comp60279_seq0 m.54339 | Presenilin-domain-containing partial | ref XP_007512557.1 | <i>Bathycoccus prasinus</i> |
| comp64356_seq0 m.57092 | Zinc transporter | ref WP_034225211.1 | <i>Actinotalea ferrariae</i> |
| comp64709_seq0 m.57251 | ATPase P | ref XP_011395511.1 | <i>Auxenochlorella protothecoides</i> |
| comp85939_seq0 m.69120 | Glycosyl transferase/Calloso synthase | ref WP_022830286.1 | <i>Cytophagales bacterium</i> |
| comp66975_seq0 m.58878 | Fucose permease | ref WP_052557275.1 | <i>Gemmata sp. IIL30</i> |
| comp59167_seq0 m.36885 | Mannose-6-phosphate isomerase | ref WP_022832831.1 | <i>Cytophagales bacterium B6</i> |
| comp69839_seq0 m.60536 | Iron/manganese transporter | gb AEW00627.1 | <i>Niastella coreensis GR2o>10</i> |
| comp76549_seq0 m.64487 | Iron/manganese transporter | ref WP_035836204.1 | <i>Cryobacterium roopkundense</i> |
| comp82989_seq0 m.65658 | Upf014 membrane protein star2 | ref WP_009735916.1 | <i>Bradyrhizobiaceae bacterium SG-6C</i> |
| comp80077_seq0 m.66613 | Diguanylate cyclase | ref WP_014280238.1 | <i>Paenibacillus terrae</i> |
| comp9596_seq0 m.8023 | Tricarboxylate transport membrane protein | | |

Table 5.4.5: A complete list of the 64 putative transporters present in both Light and Dark SCT libraries (at least one of each) with their annotation/description

| Secreted Protein Name | Annotation/Description | Top BLASTP Hit Accession | Top Hit Species |
|------------------------|---|--------------------------|---|
| comp10940_seq0 m.9287 | Unknown | - | - |
| comp11029_seq0 m.9365 | Unknown | - | - |
| comp23584_seq1 m.24185 | Unknown | - | - |
| comp13389_seq0 m.11956 | Unknown membrane component | ref WP_004718904.1 | <i>Acinetobacter guillouiae</i> |
| comp13389_seq1 m.11962 | Unknown membrane component | ref WP_004718904.1 | <i>Acinetobacter guillouiae</i> |
| comp15590_seq0 m.14704 | Hypothetical protein | ref XP_005845446.1 | <i>Chlorella variabilis</i> |
| comp19575_seq0 m.19598 | SOUL heme-binding protein | ref WP_005646909.1 | <i>Coccomyxa subellipsoidea</i> |
| comp19875_seq0 m.19979 | DBB1- and CUL4-associated factor 12-like | ref XP_005848716.1 | <i>Chlorella variabilis</i> |
| comp24544_seq0 m.25243 | Peptide ABC transporter substrate-binding protein | gb KJC48675.1 | <i>Bradyrhizobium sp. LTSP849</i> |
| comp25746_seq0 m.26477 | Chloroplast precursor | - | - |
| comp26585_seq0 m.27218 | Hypothetical protein | ref XP_005846166.1 | <i>Chlorella variabilis</i> |
| comp30431_seq0 m.30722 | Predicted protein | - | - |
| comp31345_seq0 m.31491 | Outer membrane protein | ref WP_008612716.1 | <i>Joostella marina</i> |
| comp41497_seq0 m.39631 | α -l-fucosidase | ref WP_0418861.10.1 | <i>Pedobacter sp. NL19</i> |
| comp45018_seq0 m.42444 | Dipeptidyle peptidase | ref XP_003740022.1 | <i>Metaseiulus occidentalis</i> |
| comp45978_seq0 m.43356 | Glycoside hydrolase | ref WP_049876385.1 | <i>Sorangium cellulosum</i> |
| comp48174_seq0 m.45141 | Recombinase | gb AAM49595.1 | <i>Volvox carteri f. nagariensis</i> |
| comp48206_seq0 m.45162 | Glutathione peroxidase | ref WP_022832669.1 | <i>Cytophagales bacterium B6</i> |
| comp50890_seq0 m.47213 | type I polyketide synthase | ref XP_005650993.1 | <i>Coccomyxa subellipsoidea</i> |
| comp56156_seq0 m.51365 | SNase-like nuclelease | gb AIP99476.1 | <i>Ornithobacterium rhinotracheale ORT-UMN 88</i> |
| comp57702_seq0 m.52483 | Unknown | - | - |
| comp59306_seq0 m.53753 | soma ferritin-like | gb AAN63032.1 | <i>Branchiostoma lanceolatum</i> |
| comp60645_seq0 m.54526 | Hypothetical protein | ref XP_005851273.1 | <i>Chlorella variabilis</i> |
| comp69632_seq1 m.5603 | Hypothetical protein | gb EKE16659.1 | <i>Uncultured bacterium</i> |
| comp65133_seq0 m.57547 | Raffinose synthase | ref XP_005845739.1 | <i>Chlorella variabilis</i> |

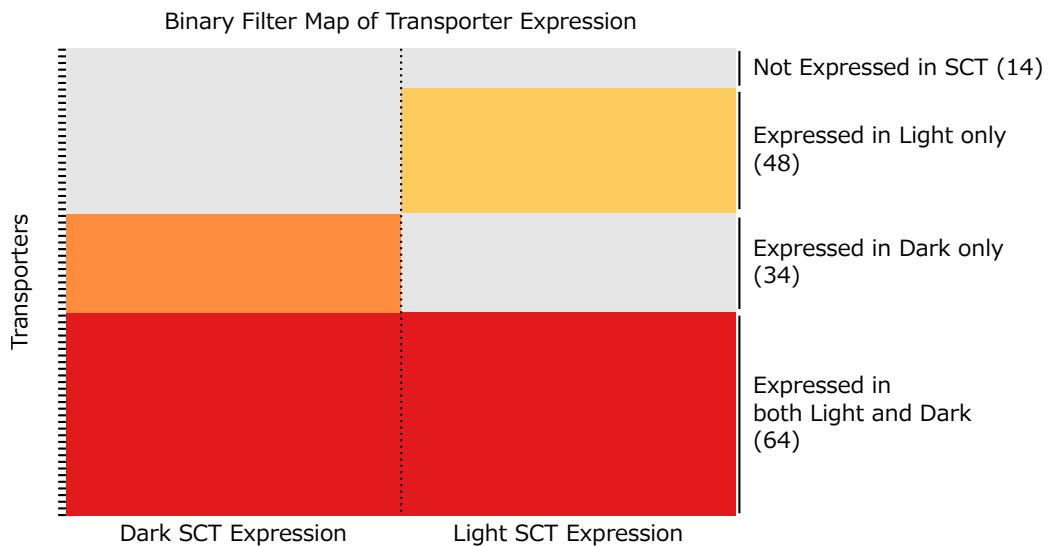


Figure 5.4.3: A binary filter heatmap that displays the 4 groups of transporter proteins identified in the *P. bursaria-M. reisseri* transcriptome. Specifically, it shows the way that transporters fall into different expression “groups” and their relative sizes. In detail, 34 transporters are expressed only in dark single cell libraries, 48 expressed only in light libraries, the 64 expressed in both and the 14 only recovered in the bulk libraries.

The secretion of a glutathione peroxidase is also intriguing. Peroxidases protect from oxidative damage therefore, this suggests there may be oxidative stress within the PV lumen. Finally, the membrane components could either be a misprediction and proteins integral to the algal outer membrane. Alternatively, they could be some of the hypothetical algal derived factors that integrate into the PV membrane (Kodama and Fujishima, 2009).

5.4.5 METABOLIC MAPS

Comparison of the *M. reisseri* metabolic map against the combined maps of the other 3 green algal datasets here (*C. variabilis* 1N, *C. variabilis* NC64A, and *C. subellipsoidea*) reveals some unique genes transcribed in *M. reisseri* that are expressed while an endosymbiont.

Specifically, aspects of carotenoid biosynthesis, a missing fumarate hydratase not present in other algae.

Additionally, there are numerous aspects of amino acid metabolism and degradation only present in *M. reisseri* compared to the other algal endosymbionts. For example, lysine degradation pathway components, and glutaminases (necessary for glutamine, D-glutamate degradation and the metabolism of alanine and aspartate).

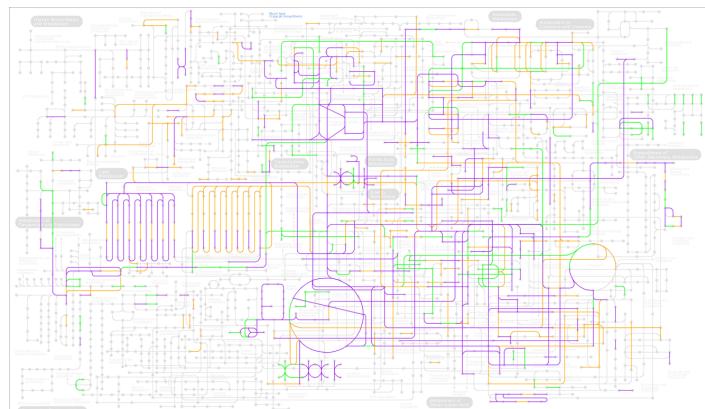
5.4.6 METABOLOMICS

5.4.6.1 GLOBAL PROFILING

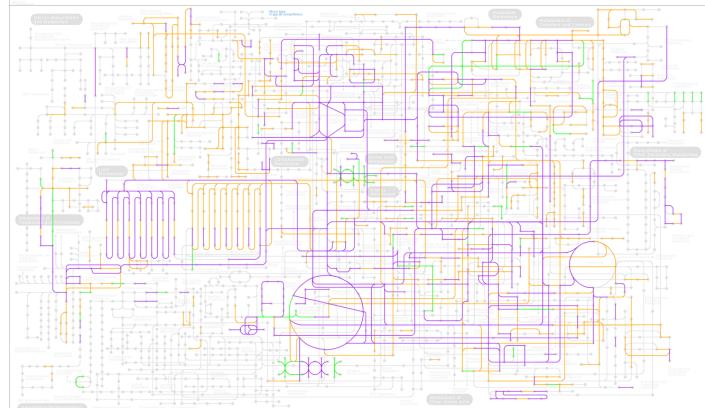
The global metabolic profiling had mixed results. Very few metabolites were identified in the GC-QTOF analysis and of those very few were the target carbohydrates. Inspection of cloud plots for this dataset (fig. 5.4.5) indicates poor calibration of GC/MS capture parameters.

3 carbohydrates were identified in this dataset (fig. 5.4.6), including a putative down-regulation of acetyl-glucose during the day. On the other hand, a putative galactose/fructose and a fucose/galactose/rhamnulose

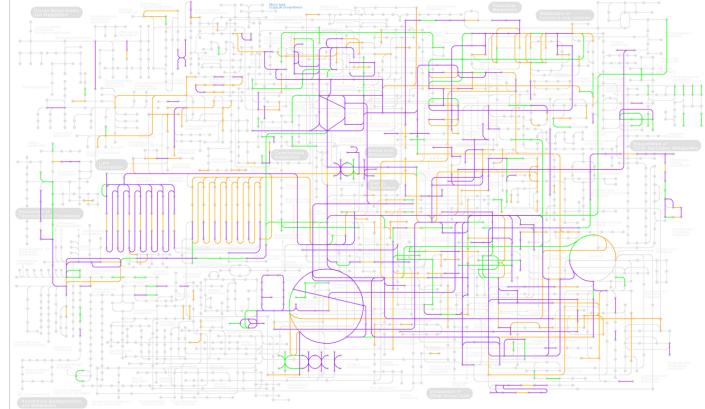
- *Micractinium reisseri* CCAP 1660/12
- *Chlorella variabilis* 1N
- Shared



- *Micractinium reisseri* CCAP 1660/12
- *Chlorella variabilis* NC64A
- Shared



- *Micractinium reisseri* CCAP 1660/12
- *Coccomyxa subellipsoidea* C-169
- Shared



- *Micractinium reisseri* CCAP 1660/12
- *Chlorella variabilis* 1N,
- *Chlorella variabilis* NC64A &
- *Coccomyxa subellipsoidea* C-169
- Shared

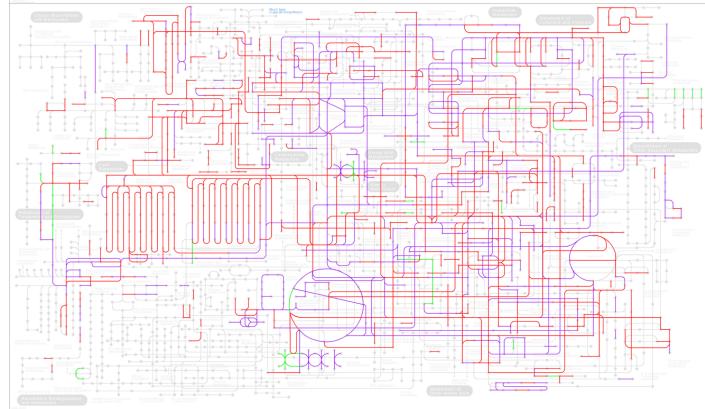


Figure 5.4.4: KEGG Maps contrasting endosymbiont bin with other algae. The bottom map highlights the aspects of *M. reisseri* putative transcriptome that are not present in the other green algal species (*C. variabilis* NC64A and 1N, and *Coccomyxa subellipsoidea* C-169).

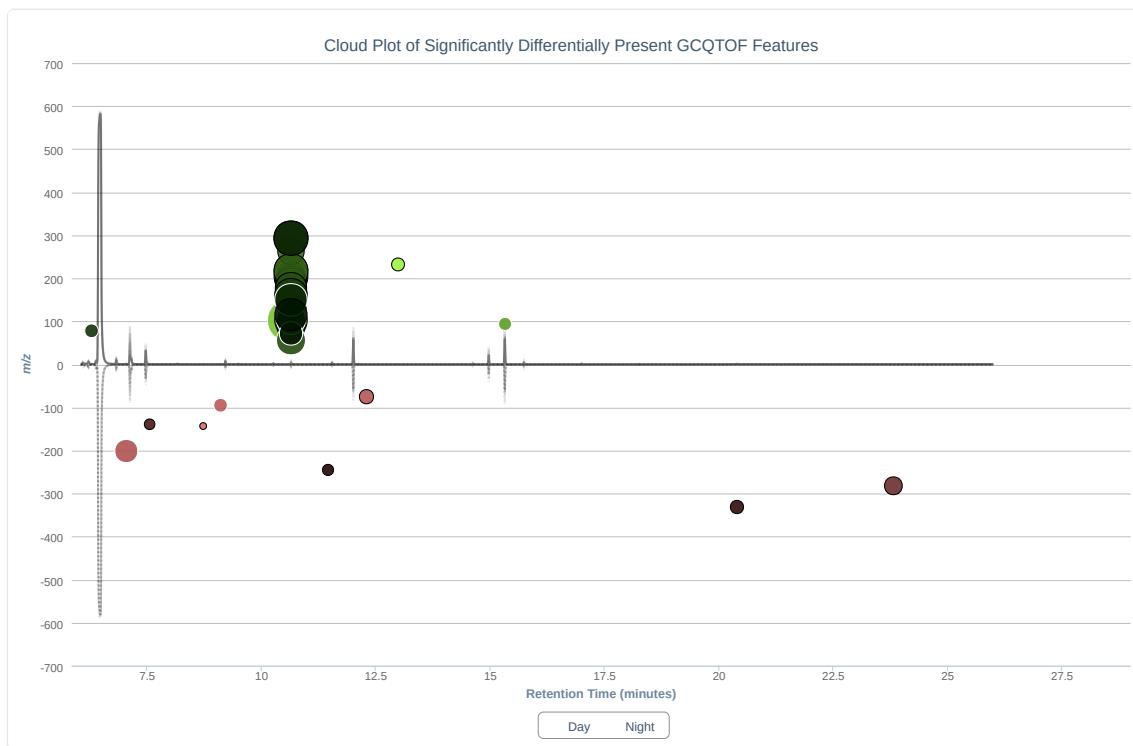


Figure 5.4.5: Cloud plot showing for the GC-QTOF analysis. The radius of a given point reflected its fold change. Data is filtered to those 50 points with significantly different expression (FDR corrected P-value of 0.01 shown by depth of colour). The poor separation of components across retention time suggests that GC could be re-calibrated to optimise the separation of these metabolites

peak demonstrated significant up-regulation during the day with fold changes between 3 and 4.

LC-QTOF seemed to be more calibrated - identifying a much greater number of metabolites (fig. 5.4.8, fig. 5.4.9, fig. 5.4.10) with better separation (fig. 5.4.7).

Several amino acids and oligopeptides were identified as significantly up and down regulated in the LC-QTOF analyses. However, the most significant peaks were that of a hugely down-regulated Raffinose (fold change of 9.7) during the day. On the other hand, the polyols Arabitol/Xylitol were up-regulated 1.52-2.24 fold and arabinose 2.7 fold during lit conditions.

5.4.6.2 TARGETED AMINO ACID ANALYSIS

The targeted quantitative analysis of the amino acids in the system between day and night also had mixed results. The majority of amino acids were not reliably recovered and calibration curves could not be fitted. Therefore, several key amino acids remain unprofiled. The relatively large concentration of Arginine as well as significant day-night differences suggests that this may be an important nitrogen source for *M. reissleri* during endosymbiosis.

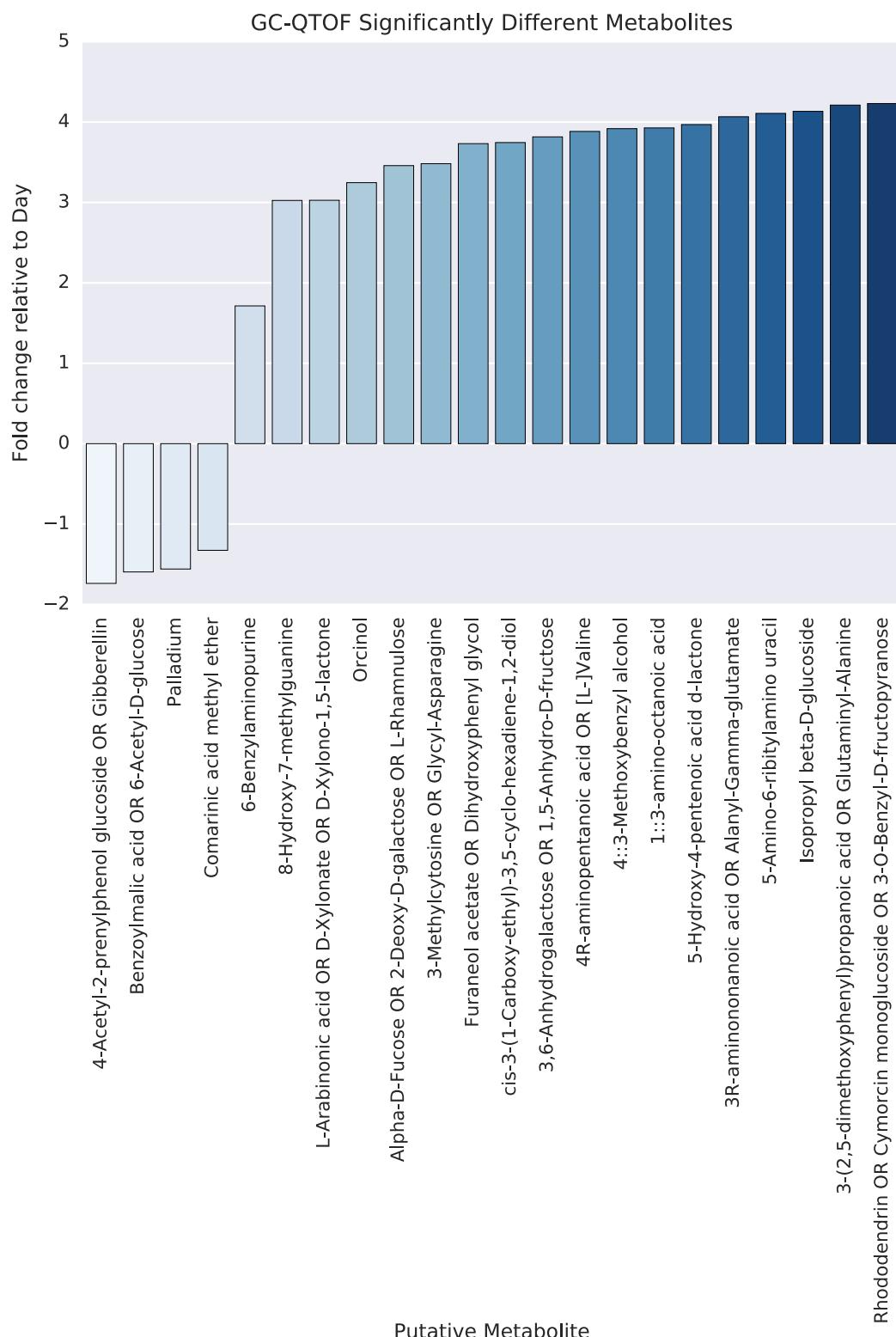


Figure 5.4.6: Plot showing the fold change of the 23 putatively identifiable significantly present metabolites from GC-QTOF. 5 peaks were discarded after inspection of the EIC. 8 were discarded as there was no sensible annotation, 14 had no annotations.

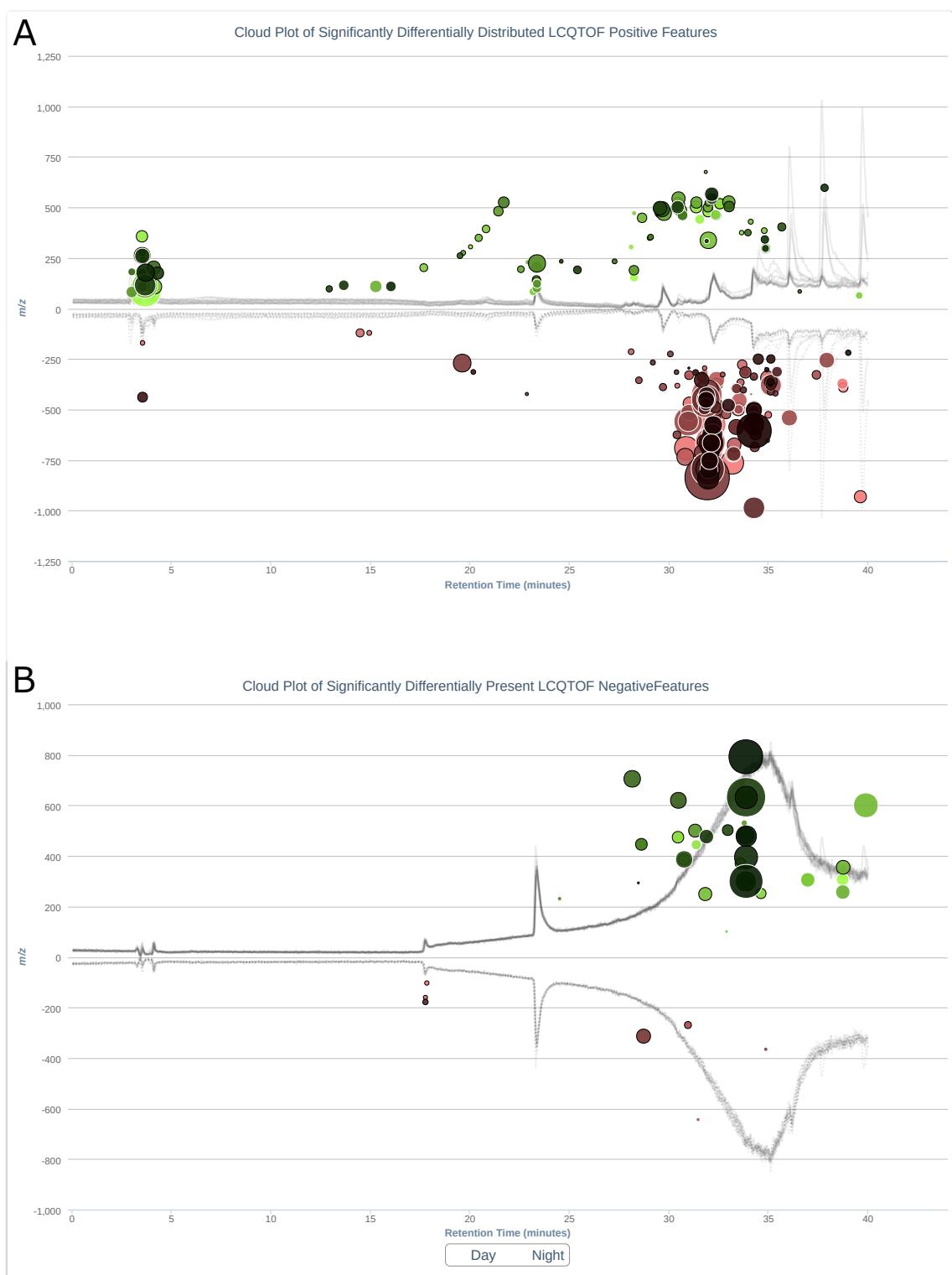


Figure 5.4.7: LC-QTOF cloud plots with A showing the positive polarity analysis and B showing the negative polarity. The radius of a given point represents its fold change, with larger radius indicating a larger fold change. Data is filtered only to include points with significantly different expression and the depth of colour indicates the P-value. Contrary to the GC-QTOF analysis LC-QTOF shows a relatively good separation of metabolites. This plot also emphasises that far more metabolites were detected and discovered differentially expressed under positive ionisation.

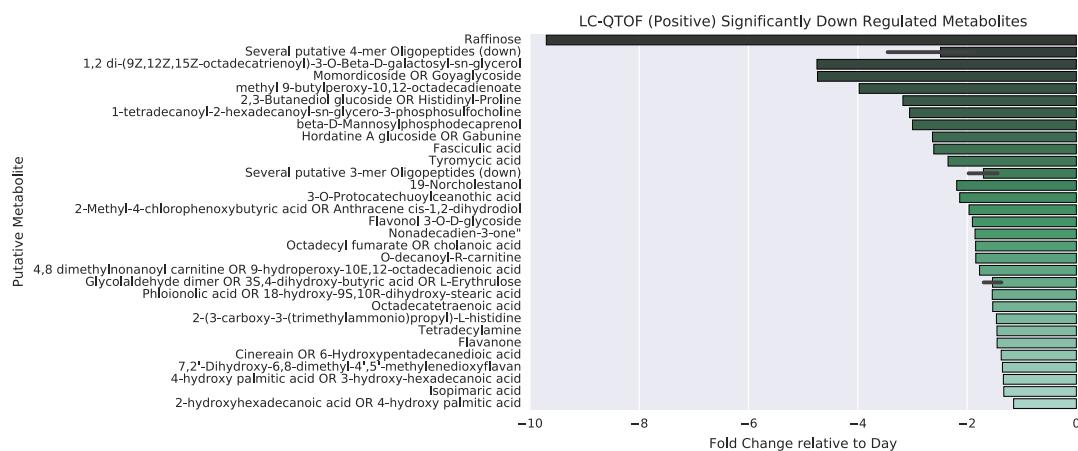


Figure 5.4.8: Of the 254 positive significantly different present metabolites, 19 were removed after manual inspection of peaks, 95 were removed due to having no METLIN hits

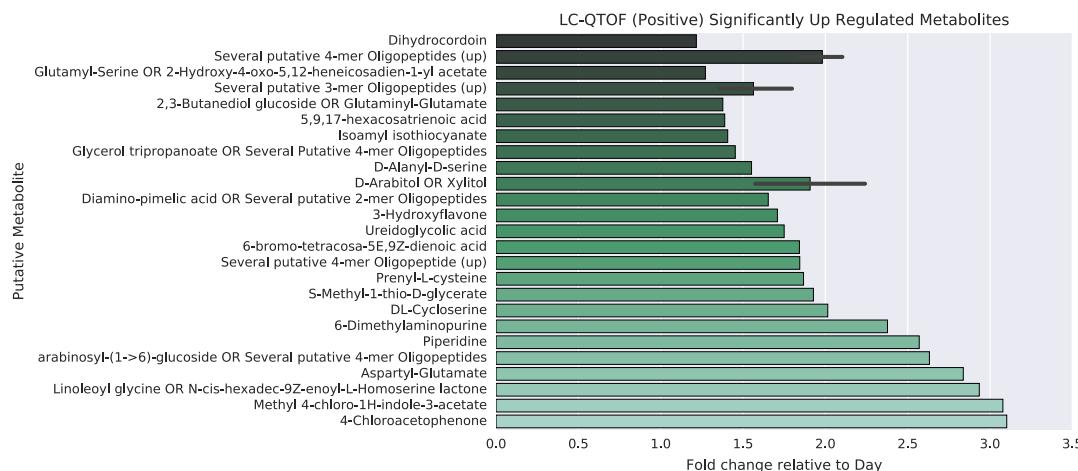


Figure 5.4.9: 254 positive significantly different present metabolites, 19 were removed after manual inspection of peaks, 95 were removed due to having no METLIN hits

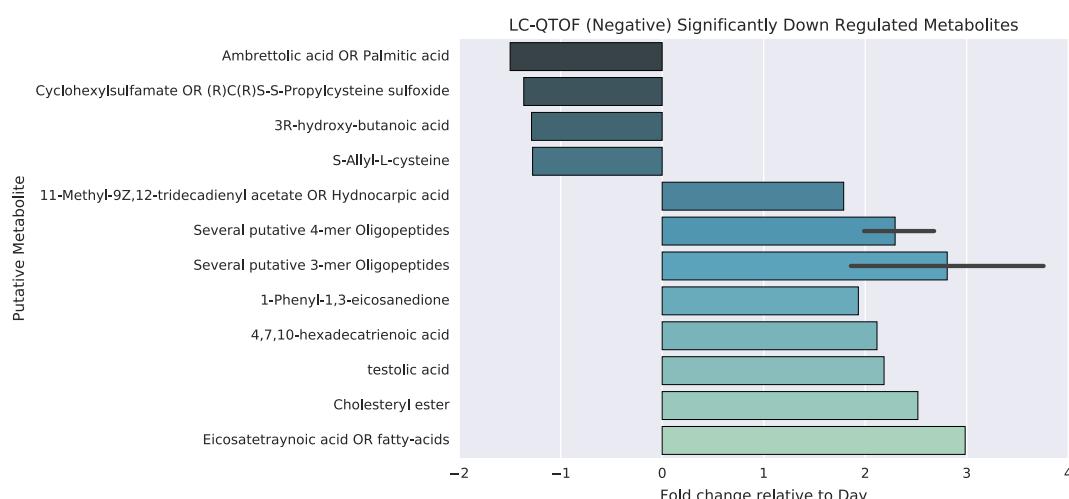


Figure 5.4.10: 43 significantly different present metabolites were identified, 3 were removed after manual inspection of peaks, 17 were removed due to having no METLIN hits

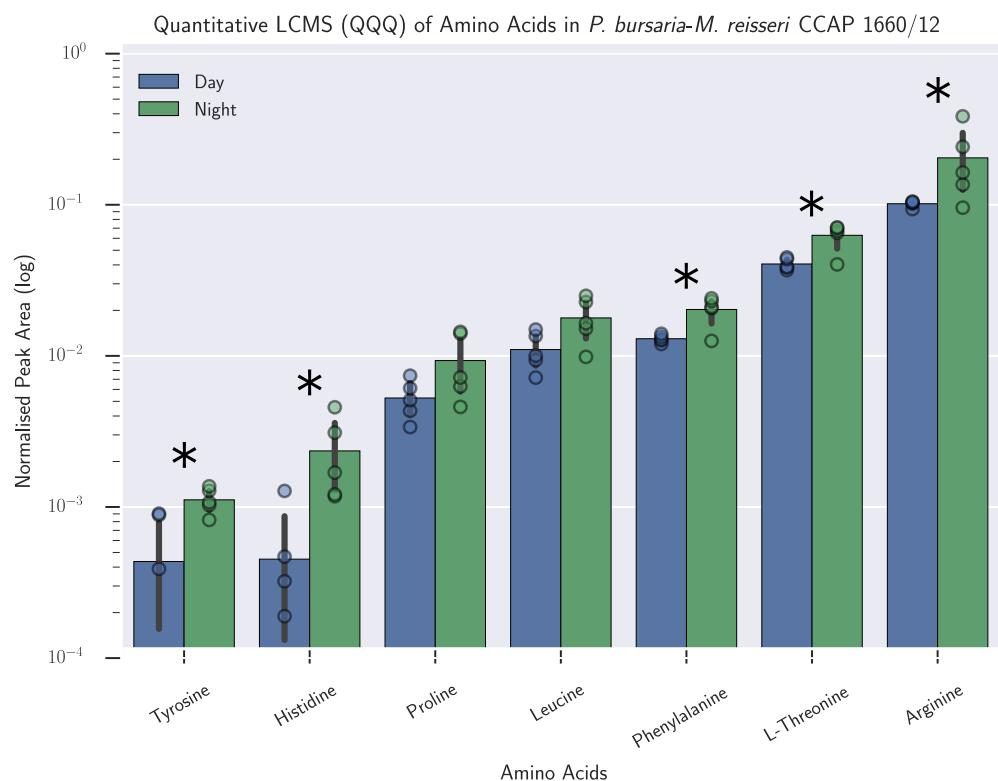


Figure 5.4.11: LC-QQQ analysis of amino acid abundances. Normalised Peak Areas Calibration was conducted using Day1 and Night1 samples at two titrations, as well as Asn-Gln-Tryptamine and Sigma AA mixes. Calibration and quantitative analysis failed for the following amino acids: Glutamic Acid, Tryptamine, Asparagine, Tryptophan, Isoleucine, Methionine, Valine, Serine, Glutamate, Glutamine, Aspartic acid, Cysteine or Lysine. Significant concentration differences between day and night (as determined via Welch's 't' are indicated by an asterisk.

5.5 DISCUSSION

5.5.1 NOVEL SUGARS IMPLICATED IN THE ENDOSYMBIOSIS

One of the key findings supported by multiple lines of evidence is the existence of roles for sugars not previously implicated in the function of *P. bursaria* - green algal endosymbioses. Specifically, arabinose, raffinose and fucose.

Fucose is identified as having a significantly higher concentration during the day and the endosymbiont expresses a fucose permease in both lit and dark conditions. Arabinose is similarly significantly at greater abundance the day however, no arabinose transporter was directly identified in the endosymbiont transcriptome. It is possible that one of the MFS group transporter identified is capable of uptake of this compound though.

Finally raffinose was both detected at significantly lower abundance during the day in the metabolomics and a raffinose synthase predicted as a secreted protein. An uptake transporter for raffinose was not identified in the endosymbiont but again it is possible that one of the MFS transporter may be capable of the uptake of this compound. Alternatively, the host could encode a transporter for raffinose uptake into the host cytoplasm. A transporter that was directly identified in the endosymbiont was that for inositol. As raffinose synthase function involves the production of inositol and raffinose and galactinol can be synthesised using inositol (Caspi et al., 2007) this suggests that inositol is either taken up by the endosymbiont or pumped into the PV lumen.

It seems clear that raffinose plays some significant role in the endosymbiosis especially as it is likely to be synthesised within the PV lumen itself due to the putatively secreted peptide. Raffinose has been associated with cold shock in *Parachlorella kessleri* (formerly *C. vulgaris*), accumulating during cold exposure and disappearing after return. Specifically, it has also been directly associated with cryoprotection of thylakoid membranes (Lineberger, 1980). Raffinose and another Raffinose Family Oligosaccharide (RFO) stachyose are also generated in gymnosperms during the cold season (Kandler and Hopf, 1982). Raffinose has also been found to inhibit growth under isosmotic conditions in a *C. vulgaris* (Setter and Greenway, 1979).

Therefore, it is non-obvious how raffinose may be involved. I present 2 hypotheses: firstly, that it may be involved in the stability and maintenance of the PV membrane due to its previously implicated role in cryoprotection of thylakoid membranes and secondly, that it may form a way in which the endosymbiont can “sequester” released carbohydrates from the host by converting them to a format the *P. bursaria* host cannot uptake. The explanation behind the significantly high concentration of raffinose at night relative to day is unclear though. This could pertain to the storage of photosynthate in the absence of active photosynthesis.

It is worth noting that the identification of other secreted proteins related to the hypothetical synthesis and metabolism of complex sugars in the PV could have been missed due to poor terminal resolution of transcripts during sequencing. This is particularly problematic for bulk RNA-seq and low terminal coverage and thus less trust-worthy data was identified in a preliminary analysis of this type of transcriptomic data. Theoretically, due to the ligation step, MDA based sc-RNAseq shouldn't have this issue. However, it seems likely the general noise and biases of MDA may have rendered a reduced positional bias as a moot issue.

5.5.2 ALTERNATIVE EXCHANGED AMINO ACIDS IN *P. BURSARIA*-*M. REISSE**R*

The failure of accurate quantification in the second round of LC-QQQ spectrometry for the majority of amino acids is problematic. However, this targeted approach still revealed useful information regarding the relative abundance of amino.

Particularly, both the high concentration of arginine as well as its significantly differential abundance between day and night indicates that this amino acid may well form a major component of host provided nitrogen for *M. reisseri*. This is in concordance with previous findings suggesting the importance of this amino acid in the *C. variabilis* endosymbiosis ([Kato et al., 2006](#)).

Despite not have been implicated in previous analyses as one of the key amino acid nitrogen sources, the identification of both high quantities and differential abundant threonine and phenylalanine suggests that these amino acids may play a role in the host-endosymbiont barter system of *P. bursaria*-*M. reisseri*.

Additionally, the unique presence of lysine, glutamine and D-glutamate degradation pathways in *M. reisseri* relative to other green algae suggests that these amino acids also comprise an element of the host-derived nitrogen supply.

The difference in amino acids abundances for several of those fitted as well as differential numbers of reads mapping to putative amino acids transporters indicates light-dependent amino acid uptake in the endosymbiont.

This alternative behaviour suggests that the feeding experiment results by ([Kato et al., 2006](#)) and ([Kato and Imamura, 2009b](#)) need to be re-evaluated for *M. reisseri*. This also adds further support to the diversity of endosymbiotic relationships between *P. bursaria* and its various algal endosymbionts.

Unfortunately, a failure to accurately quantify and calibrate for several amino acids means this targeted metabolomic analysis is incomplete. Of particular interest, is the remaining 5 amino acids implicated in *C. variabilis* F36-ZK's endosymbiosis.

One other observation is the potential supply of oligopeptides to the endosymbiont. The presence of a partial oligopeptide transporter combined with the high concentrations of various 3- and 4-mer oligopeptides suggests that previous studies focusing exclusively on endosymbiont utilisation and uptake of individual amino acids (e.g. those reviewed in ([Kato and Imamura, 2009b](#))) may have missed on the role of these in host-endosymbiont nitrogen flux.

In terms of other nitrogen sources, i.e. nitrate and nitrite: NR and NiR are both present in the *M. reisseri* binned transcriptome. This supports findings that *M. reisseri* can utilise nitrate and nitrite it is possible that these are non-functional like the mutants in *C. variabilis* NC64A and F36-ZK.

5.5.3 POTENTIALLY MISSING TRANSPORTERS AND SECRETED PROTEINS

The identification and analysis of secreted and transporter proteins, as well as the metabolic mapping are fundamentally reliant on the quality and completion of the host and endosymbiont binned transcriptomes. Transcripts may be missing from these bins either due to failure to assemble, cryptic MDA biases, or erroneous binning into bins other than host or endosymbiont.

It is a cause for concern that there are a number of endosymbiont secreted and transporter proteins which have top BLASTP hits against bacterial species. On inspection many of these do have other hits to green algal species at slightly lower expectation therefore, binning could be working as intended. However, bacterial contamination of the endosymbiont bin is a very real possibility. Further studies should confirm the identity of these proteins using manual in-depth phylogenetics instead of the high-throughput and potentially error prone method using in transcript binning.

There is evidence to the incompleteness of the binned transcriptomes in the large disparity in predicted transporter set sizes between the genomes and transcriptome based datasets. However, we are not necessarily interested in all the transporters and secreted proteins that the endosymbiont is capable of producing but merely those that it is producing while an endosymbiont. Therefore, the only major group of factors that are both involved in endosymbiosis and systematically absent from these binned transcripts are those of effectors related to the establishment of endosymbiosis that are not expressed during the rest of the endosymbiotic relationship.

One set of proteins in which erroneous binning may be particularly problematic is that of proteins which have recently undergone endosymbiotic gene transfer (EGT) between endosymbiont and host or that have been horizontally acquired from other sources. In the case of the former case, this is well observed phenomenon that has resulted in the eventual loss of the endosymbiotic in the majority of algal secondary endosymbiotic organelles as genes are transferred to the host nucleus (Keeling and Palmer, 2008; Archibald, 2005; Keeling, 2004). It is unknown and difficult to determine to what extent the unusual nuclear dimorphism and germline sequestration of the host will effect this form of transfer. However, hypothetically this should present a barrier to such transfers. As some *M. reisseri* and *Chlorella* endosymbionts have been demonstrated as capable of free-living and metabolic co-dependence has putatively not become fixed it is unlikely that EGT has occurred between host and endosymbiont as extensively as that observed in established photosynthetic organelles. Fortunately, the binning method used means that while some peptides may have been falsely assigned to wrong bin all “host” and “endosymbiont” ORFs that were not either so novel they lacked any homology to known proteins or were recently acquired from bacteria were still included in this analysis just not necessarily attributed to the correct partner.

As for the latter case of HGT from other sources, this will lead to the misclassification of proteins into the “food” or “unknown” bin and thus their discard. This is potentially problematic as there is evidence for bacterially acquired hexose-phosphate transporters playing a key role in the establishment of primary plastid endosymbiosis (Price et al., 2012; Karkar et al., 2015).

Ideally, future work could expand the component analyses over the “unknown” and “food” binned sequences in combination with synteny analyses using genomic sequences to investigate and identify potential horizontally acquired transporters that may play a role.

Another issue with the binning approach used is the possibility of totally novel transporter (and other proteins) not being classified due to the dependence of the binning on homology to known sequences. Therefore, totally novel proteins would not have been identified as “host” or “endosymbiont”. Unfortunately, this problem could only properly be resolved with a genome sequence for both host and endosymbiont which was outside the scope of this analysis.

Finally, there are 2 additional difficulties specific to transporter and secretome prediction. In terms of secreted proteins, without knowledge of *P. bursaria*'s intracellular trafficking system it is not possible to easily infer which host peptides are secreted into the PV. For this reason, analysis of secreted proteins focused on the endosymbiont bin as the secretion signal are generally better conserved and established.

In the case of transporters the risk of false positives where sensors are misidentified as transporters is fairly high. Only a minimal (as few as a single amino acid) divergence is sufficient to convert a transporter protein to non-transporting sensor proteins (Lalonde et al., 1999; Bianchi and Díez-Sampedro, 2010). However, sensor proteins are likely to play important roles in the function of this endosymbiosis so accidental identification may not be a major issue but it does mean transporter activity needs verified for all predictions.

Prediction methods used to identify transporter and secreted proteins could also be further improved by more careful application of state-of-the-art classification algorithms. Prediction of protein secretion could theoretically be improved using approaches like recursive neural network models, specifically long-short term memory networks (Hochreiter and Schmidhuber, 1997; Greff et al., 2015). These are well suited to arbitrary length sequence data so wouldn't require work-arounds to accommodate variable length signal peptides. They are also capable of representation learning so the identification of sequence features that best predict localisation would be unnecessary. Alternatively, existing prediction tools could be combined in a more sophisticated way than the conservative consensus ensemble used in this analysis. For example, the various predictors could be combined using Bayesian model combination (Monteith et al., 2011).

However, even with these improved predictions, it is still best to consider all analyses of these transcriptomes as proof of presence but not proof of absence of any component. Furthermore, any identified protein should be validated individually using immunolocalisation and (q)PCR based analyses.

5.5.4 METABOLOMICS SHOWS PROMISE

The pilot application of metabolomics demonstrated mixed results. There was poor performance of GC/MS with a failure to comprehensively profile carbohydrates. Several compounds strongly implicated in the endosymbiosis being undetected e.g. maltose, and glucose. This was potentially due to the miscalibration of gas chromatography leading to poor separation of components.

While LC/MS analyses did prove relatively successful a careful validation of metabolites of interest using multi-reaction monitoring and tandem spectrometry would be necessary to make firm predictions. Additionally, advanced novel techniques such as nanoscale secondary ion mass spectrometry combined with microscopy and isotope labelling could theoretically allow direct analysis of metabolites present in the PV (Kopp et al., 2015; Legin et al., 2014).

The targeted quantitative LC/MS of amino acids needs further optimisation and re-running due to the inability to fit calibration curves to the majority of amino acids. Despite this, the technique did prove effective at identifying a potential role for several amino acids not previously implicated in this endosymbiosis.

Finally, one more improvement to the metabolomics analyses would include more intelligent hypothesis testing than corrected unequal variance 't'-test could be used such as Kurschke's Bayesian BEST algorithm (Kruschke,

2013). This also has the advantage of a Bayesian inference which can be made robust to multiple comparisons without need for extensive correction procedures by using standard multi-level approaches (Gelman et al., 2009).

5.6 CONCLUSION

This analysis of host-endosymbiont metabolic integration has lead to some promising results. Namely, discovering quantitative data supporting the mechanism by which the host likely provides a nitrogen source to the endosymbiont. Specifically, a novel group of amino acids may well be used in *M. reisseri* endosymbiosis: lysine, d-glutamate, threonine, and phenylalanine. As well as the previously implicated arginine and glutamine (Kato and Imamura, 2009b). Additionally, potential novel roles for carbohydrates previously not associated with *P. bursaria* endosymbioses, specifically fucose, arabinose and raffinose have been identified. Unfortunately, poor resolution and identification of carbohydrates in GC/QTOF prevents a thorough analysis of endosymbiotic carbohydrate metabolism. Finally, *M. reisseri* appears not to express elements of fatty acid degradation present in the other algal endosymbionts.

However, further work is needed to confirm both the role and localisation of these compounds and their facilitators.

Ultimately, the key finding from this analysis is that *M. reisseri* exhibit a range of adaptations to endosymbiosis that are distinct from previously studied algal strains such as *C. variabilis*.

In biology, nothing is clear, everything is too complicated, everything is a mess, and just when you think you understand something, you peel off a layer and find deeper complications beneath. Nature is anything but simple.

- Richard Preston: The Hot Zone

6

RNAi in *P. bursaria*

6.1 INTRODUCTION

6.1.1 RNAi

Post-transcriptional gene silencing (PTGS) is a highly useful experimental technique in reverse genetic analyses. The most widely used PTGS experimental method is that of RNA-mediated interference (RNAi) of gene expression (Fire et al., 1998). It has extensively used in the study of model eukaryotic organisms (Morf et al., 2013; Batista and Marques, 2011; Matthew, 2004; Ketting, 2011; Chang et al., 2012).

RNAi covers a whole set of evolutionarily conserved systems across the eukaryotes with various mechanisms of action in which the expression of particular transcripts are regulated via several classes of transcribed small non-coding RNA (ncRNA) such as short-interfering (siRNA), micro (miRNA) and Piwi-interacting (piRNAs) (Carthew and Sontheimer, 2009).

These systems likely originated as a form of defence against viruses and transposons (Waterhouse et al., 2001; Buchon and Vaury, 2006) and were present in some form in the last universal eukaryotic ancestor (LECA) (Cerutti and Casas-Mollano, 2006; Shabalina and Koonin, 2008). Many eukaryotes utilise these small RNA-mediated gene silencing pathways in the regulation of their own cell expression patterns (Wu and Belasco, 2008). Despite its ancestral nature there has been considerable diversification of both this process, its function and mechanism (Ketting, 2011). Indeed, even within the same organism, different points of the life cycle may use different RNAi

systems (Flemr et al., 2013).

Generally RNAi pathways involve the generation of 21-28nt siRNAs from some form of RNA precursor such as dsRNA (although ssRNA systems exist) via the function of the RNAase III Dicer (Bernstein et al., 2001) or a related protein. These short RNAs are then bound by Argonaute proteins which act alone or as part of a complex to silence the expression of sequences homologous to the siRNA (Ketting, 2011). This silencing isn't just limited to the post-transcriptional endonucleolytic degradation of mRNA transcripts but can also involve transcriptional inhibition and DNA elimination (Marker et al., 2014). The one unifying element of all discovered RNAi pathways is that of the central role argonaute (AGO) proteins play (Ketting, 2011). They are formed of two subclasses: the Ago and Piwi subfamilies (Peters and Meister, 2007) with a range of functions and complex-forming behaviours (Ender and Meister, 2010). The magnitude of the silencing response is occasionally amplified by the generation of more copies of the trigger dsRNA by RNA-dependent RNA polymerases (RdRPs) (Arp et al., 2007). On the other hand RdRPs can also sometimes directly generate the siRNAs (Aoki et al., 2007; Ketting, 2011). The last universal eukaryotic ancestor (LECA) likely contained at least one Ago and Piwi family Argonaute protein, a Dicer and an RdRP (Cerutti and Casas-Mollano, 2006).

The other main form of RNAi system present in eukaryotes is that of miRNA based systems. These are differentiated by miRNAs being encoded by dedicated genes and displaying partial complementarity to their targets whereas siRNAs are generated from exogenous dsRNAs (i.e. environmental dsRNA from viral infection or phagocytosed bacteria (Whangbo and Hunter, 2008)) or transgenes as described above and involve full or near full complementarity (Shabalina and Koonin, 2008).

On top of this, there are piRNA based systems, which are involved in germline based transposon silencing (Iwasaki et al., 2015), and the ciliate specific scan RNA (scRNA) system. This is involved in the elimination of internal eliminated sequences (IESs) during macronuclear (MAC) regeneration (Mochizuki and Gorovsky, 2004; Chalker et al., 2013).

Experimentally, the existence and function of these systems permits a researcher to introduce dsRNA homologous to an RNA transcript of interest and trigger targeted cell-wide RNAi of that transcript. Unfortunately, there are also several problems with RNAi as a general method. Many organisms lack active RNAi systems (although such systems can occasionally be induced (Alibu et al., 2005)). On top of this, RNAi requires accurate sequence data to design the precursors therefore necessitates some form of sequencing. The main difficulty, however, is that of off-target effects. These are caused when a provided sRNA precursor induces RNAi in more than just the target transcript. These can lead to enigmatic phenotypic outcomes that are then falsely attributed to the initial target. Avoidance of off-target effects requires a complete genome and/or transcriptome to check a prospective siRNA against during the design stage. This further increases the epistemological burden of attempting RNAi in a novel system.

Finally, RNAi does not necessarily induce total silencing of a given transcript and low-levels of transcription may still occur. This, conceivably, can be sufficient to maintain the non-knock down phenotype. This allows a researcher to falsely conclude a non-relationship between a given transcript and phenotype.

| Pathway | Component | Function |
|-------------------------------|-----------|-------------------------|
| transgene-induced siRNA | Rdr3 | RdRP |
| both pathways | Ptiwi14 | Piwi |
| | Rdr2 | RdRP |
| | Dcr1 | Dicer |
| | Ptiwi13 | Piwi |
| exogenous dsRNA-induced siRNA | Cid2 | Nucleotidyl transferase |
| | Rdr1 | RdRP |
| | Cid1 | Nucleotidyl transferase |
| | Ptiwi12 | Piwi |
| | Ptiwi15 | Piwi |
| | Pds1 | Import of dsRNA? |

Table 6.1.1: Summary of the components identified as necessary to the function of both primary siRNA RNAi pathways in *P. tetaurelia* as identified by forward genetic screens in (Marker et al., 2014)

6.1.2 RNAI IN PARAMECIUM

In addition to the ciliate specific scnRNA system siRNA based pathways have been discovered in the two principal ciliate model organisms: *Tetrahymena thermophila* (Collins and Lee, 2006; Yao and Chao, 2005) and *Paramecium tetaurelia* (Galvani et al., 2001; Galvani and Sperling, 2002). There are 2 established methods for inducing RNAi in *Paramecium tetaurelia*: microinjection and transformation of the MAC with high-copy transgenes lacking 3' untranslated region (UTR) (Galvani et al., 2001) and the introduction of dsRNA by either microinjection or feeding using transformed dsRNA expressing bacteria (Galvani and Sperling, 2002).

In the transgene pathway, the 3' truncation leads to the production of aberrant sense and antisense transcripts (Galvani et al., 2001; Marker et al., 2010; Beisson et al., 2010a). Based on the identified required components (see table 6.1.1), aberrant transcripts are processed by a Dicer protein (Dcr1) (Lepere et al., 2009) and a putative RdRP complex formed of an RdRP (Rdr2) and a nucleotidyl transferase (Cid2) (Marker et al., 2014) into 23nt siRNA (Lepere et al., 2009). A putatively non-catalytic RdRP (Rdr3) also plays an undefined role in the generation of primary (1°) siRNAs from transgene pre-cursors (Marker et al., 2010, 2014). Finally, two Argonaute Piwi proteins (Ptiwi13 and Ptiwi14) (Bouhouche et al., 2011) are involved in targeting post-transcriptional silencing via mRNA cleavage (Bouhouche et al., 2011; Marker et al., 2014).

Alternatively, the exogenous dsRNA pathway can be induced by either microinjection directly into the MAC (for a transient 48 hour long silencing) or by continued feeding with a bacterially experimentally modified to generate dsRNA. Typically, this involves an *E. coli* with an IPTG-inducible T7 polymerase and deficiency for RNase III transformed with a plasmid containing a T7 promoter and the sequence homologous to the target transcript (Fire et al., 1998; Timmons et al., 2001; Galvani and Sperling, 2002). Importantly, there is some evidence that this pathway also is activated at low levels by ssRNA from normal food bacteria (Carradec et al., 2015).

Based on the components that have been identified as necessary for the exogenous dsRNA pathway to function (see table 6.1.1). RNA precursors are processed by Dcr1 (Lepere et al., 2009), and then two hypothetical RdRC (Cid1-Rdr1, Cid2-Rdr2) (Marker et al., 2010, 2014) are involved in the generation of 1° siRNA. The second RdRC (or specifically Rdr2) is also involved in the generation of low levels of secondary (2°) siRNA which spread along the full length of target mRNA (i.e. 3'-to-5' and 5'-to-3' transitivity) in primarily antisense form (Carradec et al.,

2015). These 2° siRNAs don't appear to play a significant role in silencing themselves (contrary to similar systems in *C. elegans* where they form the principal targeter of silencing (Sijen et al., 2007; Pak and Fire, 2007)) (Carradec et al., 2015). 3 Piwis play a role in targeting silencing. Ptiwi13 hypothetically loads the 1° siRNA and targets cleavage of cytoplasmic mRNA (Bouhouche et al., 2011), while Ptiwi12 and Ptiwi15, based on their homology to nuclear Piwi proteins, (Marker et al., 2014; Carradec et al., 2015; Bouhouche et al., 2011) may be involved with 2° siRNA communication with the MAC (Carradec et al., 2015). One final protein is necessary for the function of feeding based dsRNA-induced RNAi is an uncharacterised novel *P. tetaurelia* complex protein (Pds1) (Marker et al., 2014). It has been hypothesised that Pds1 may play a role in the export of RNA from the food vacuole (Carradec et al., 2015). Therefore, theoretically microinjected dsRNA should induce RNAi even in the absence of this protein as the microinjection circumvents the need for Pds1 facilitated dsRNA import.

Many of these components are a product of the 3 whole genome duplication events in the evolution of the *Paramecium* clade (McGrath et al., 2014). As *P. bursaria* shares only the first *Paramecium* clade whole genome duplication event with *P. tetaurelia* it is expected it should contain the RNAi components identified as belonging to WGD1 (or have secondarily lost them) (McGrath et al., 2014). This is believed to include a single RdRP gene, 6 Piwi genes, and 2 Dicer genes (Marker et al., 2014).

If it is possible to experimentally induce RNAi in *P. bursaria* SW1 (from CCAP 1660/12 culture) specific hypotheses as to the necessity of hypothetically important endosymbiotic components can be tested. For example, what is the effect on endosymbiosis of the inhibition of certain host-derived transporters.

Similarly, what components of core *P. tetaurelia* RNAi pathways can be identified in the *P. bursaria* SW1 (see Chapter 4) and *P. bursaria* Yad1g transcriptomes (see Chapter 5)? Does *P. bursaria* express these components during endosymbiosis? If they don't is there evidence of them in the partial *P. bursaria* SW1 genome (see Chapter 3)?

6.1.3 RNAi "CROSS-TALK"

Evidence has emerged of the role RNAi plays in numerous host-pathogen (Nowara et al., 2010; LaMonte et al., 2012; Weiberg et al., 2013; Buck et al., 2014) and host-symbiont (Helber et al., 2011; Koch et al., 2013) relationships. This has led some authors to suggest that siRNA and RNAi mechanisms can form communication systems between diverse organisms and even across domains (Liang et al., 2013; Knip et al., 2014; Weiberg et al., 2015).

The evidence of natural food bacteria ssRNA induced RNAi "cross-talk" in *P. tetaurelia* (Carradec et al., 2015) implicates that this process may also take place in *P. bursaria*. In addition to also being a serial phagotroph *P. bursaria* acts as a host to numerous bacterial and green algal endosymbionts. Therefore, it is not inconceivable that such a mechanism of cross-talk many play role in these endosymbioses.

Therefore, it may be informative to investigate the quantity and targets of potential cross-talk between host and endosymbiont in terms of "collisions" i.e. matching 23nt RNA strings between host and endosymbiont transcripts bins. Contextualising these values across the diversity of the tree of life is important. "Collision" levels have implications for the regulation and expression of exogenous dsRNA RNAi pathways by the host.

| Gene | Function | RNAi phenotype in <i>P. tetaurelia</i> | Vector Design | Reference |
|--------------|----------------------------|--|--|------------------------------|
| <i>epi2</i> | Epiplasmin | "Monstrous" cells | 500bp via <i>PstI</i> and <i>HindIII</i> | (Damaj et al., 2009) |
| NSF | Membrane fusion factor | Lethal | 500bp via <i>PstI</i> and <i>HindIII</i> | (Galvani and Sperling, 2002) |
| pTMB_422c | Binding protein | Lethal | 500bp via <i>PstI</i> and <i>HindIII</i> | (Nowack et al., 2011) |
| <i>bug22</i> | Basal body/ciliary protein | Slow swimming and death | 313bp via <i>XbaI</i> and <i>HindIII</i> | (Laligne et al., 2010) |
| BBS7 | Ciliary ion transport | Fewer, shorter ciliar | 486bp via <i>XbaI</i> and <i>HindIII</i> | (Valentine et al., 2012) |
| PGM | PGM endonuclease | Post-autogamous cells unable to resume normal growth | 500bp via <i>PstI</i> and <i>HindIII</i> | (Baudry et al., 2009) |

Table 6.3.1: Details of RNAi vectors used in dsRNA experiments. All constructs were cloned into a L4440 vector and used an Ampicillin resistance markers.

6.2 AIMS

The goal of this chapter is to investigate both the practical and theoretical utility of RNAi systems in *P. bursaria*. Specifically:

- Is *P. bursaria* capable of microinjection or feeding based exogenous dsRNA siRNAi?
- What components, previously identified as necessary, for these pathways are present and expressed in *P. bursaria*?
- To what degree could potential RNAi "cross-talk" occur in *P. bursaria* is this elevated compared to what might be faced in *Paramecium* species without eukaryotic endosymbionts?

6.3 METHODS

6.3.1 RNAI CONSTRUCTS

All RNAi methods were based on previously published protocols, specifically (Galvani et al., 2001; Galvani and Sperling, 2002; Beisson et al., 2010b).

Six different constructs were created featuring genes whose knock-down induces known phenotypes in *P. tetaurelia* (see table 6.3.1). All inserts were designed in the same manner, *P. tetaurelia* sequences were taken from the *P. tetaurelia* genome. These were then BLASTN against the entire unbinned *P. bursaria-M. reissieri* SW1-ZK transcriptome and designed based on the *P. bursaria* sequence. Each insert was cloned into an L4440 vector featuring two convergent T7 promoters and an ampicillin resistance marker.

6.3.1.1 RNAI FEEDING

For feeding experiments the vectors were transformed into *E. coli* HT115-DE3. This strain is deficient for RNase III and features an IPTG inducible T7 polymerase under the control of a Plac promoter. Method used was as described in (Beisson et al., 2010b). RT-PCR using standard methods was conducted on the transformed cultures to confirm expression of the dsRNA.

Bacterial precultures were started using a single colony picked from an LB plate containing $50\mu\text{gml}^{-1}$ ampicillin and $12.5\mu\text{gml}^{-1}$ tetracycline. This picked colony was grown overnight in LB medium with the same antibiotics. The overnight culture was then diluted 50 fold and grown with shaking at 37°C up to an OD_{600} of 0.4 to 0.6. IPTG was then added at a concentration of 0.4mM and shaken for 3 hours at 37°C . 30ml of this culture was

centrifuged for 2 minutes ($3100 \times g$), then the supernatant removed and the pellet washed twice in *Paramecium* growth medium. The pellet was resuspended in *Paramecium* medium with 0.4mM IPTG, $100\mu\text{gml}^{-1}$ ampicillin and adjusted to a final OD_{600} of 0.1. $1\mu\text{l}$ of beta-sitosterol (at 4mgml^{-1} in ethanol) was added to each 5ml of medium.

For the actual feeding, 10ml *P. bursaria* CCAP 1660/12 culture were centrifuged at $800x g$ for 10 minutes and re-suspended in 1ml of supernatant. 9ml of the induced bacterised media was then added. The sample was then incubated in a tissue culture flask at 27°C . Feeding was repeated for each day of analysis.

6.3.1.2 RNAI MICROINJECTION

Microinjection used the same protocol as described in (Beisson et al., 2010a) but only tested the PGM construct. Briefly, the circular plasmid is linearised using a unique restriction site, and purified using phenol:ethanol extraction and a purification column. It is then dissolved in H_2O at a minimum concentration of 5mgml^{-1} . Cells were washed twice by picking with a micropipette in Dryl-BSA wells before being placed into individual droplets on a glass coverslip and covered in paraffin oil. The coverslip was then placed onto a microscopy stage and a microinjector used under $10X$ magnification to inject linearised construct directly into the macronucleus.

Microinjection controls were also conducted in which GFP was injected into the MAC and the fluorescence observed. Successful microinjection would feature fluorescence localised to the MAC.

6.3.2 ANALYSIS OF RNAI PATHWAY

6.3.2.1 SURVEY FOR RNAI COMPONENTS IN *P. BURSARIA*

Using the canonical seed sequences identified in *P. tetaurelia* by (Marker et al., 2014) (see table 6.3.2) the entire assembled *P. bursaria*-*M. reisseri* SW1-ZK transcriptome predicted ciliate encoding peptides and *P. bursaria*-*C. variabilis* Yad1g1N transcriptome predicted ciliate encoding peptides with BLASTP and a minimum expectation of $1e^{-5}$.

Finally, if a component could not be identified in the transcriptomes it was additionally searched for in the assembled *P. bursaria*-*M. reisseri* SW1-ZK genomic contigs (over 500bp) using tBLASTx with a minimum expectation $1e^{-5}$. This would theoretically allow the identification of present but non-expressed components.

Additionally, the other sequenced *Paramecium* genomes were searched using BLASTP via ParamediumDB (Arnaiz et al., 2007; Arnaiz and Sperling, 2011b). Specifically, *P. caudatum* (McGrath et al., 2014), *P. biaurelia*, *P. primaurelia*, *P. sexaurelia* and *P. multimicronucleatum*. Finally, *T. thermophila* (Eisen et al., 2006) and *Oxytricha trifallax* (Swart et al., 2013) predicted proteins were searched to form outgroups during phylogenetic analysis.

6.3.2.2 PHYLOGENETIC ANALYSIS OF RNAI PATHWAY

Peptide sequences were aligned using MAFFT (Katoh et al., 2002) and manually masked in Seaview (Gouy et al., 2010). Sequences that were too divergent to align were removed, or in the case of the RdRPs: the alignment

| Gene | P. tetaurelia Accession | Length |
|-------------|--------------------------------|---------------|
| Rdr1 | PTETG8500012001 | 4319 |
| Rdr2 | GSPATG00036857001 | 4162 |
| Rdr3 | GSPATG00006401001 | 3292 |
| Cid1 | PTETG9100013001 | 1051 |
| Cid2 | PTETG13400003001 | 1083 |
| Pds1 | PTETG600032001 | 2084 |
| Dcr1 | GSPATG00021751001 | 5394 |
| Ptiwi12 | GSPATG00001709001 | 2315 |
| Ptiwi13 | PTETG4800007001 | 2483 |
| Ptiwi14 | PTETG16300003001 | 2428 |
| Ptiwi15 | GSPATG00005370001 | 2315 |

Table 6.3.2: Table of the RNAi pathway components identified by (Marker et al., 2014). Includes their “canonical” accession in the *P. tetaurelia* MAC genome.

of Rdr1 and Rdr2 was split from that of Rdr3 to form two separate alignments. Similarly, sequences that were identical to one another were removed at this stage.

Substitution models were fitted using ProtTest3 on the basis of the Bayesian Information Criterion (BIC) (Darriba et al., 2011). Phylogenies were then generated using RAxML with 1000 non-rapid bootstraps and Mr-Bayes with 2 runs of 4 MCMCMC chains run for 2,000,000 generations or until convergence. MCMC convergence was checked and burn-in determined using Tracer (Rambaut and Drummond, 2007). Sequences forming long branches were removed after inspection of these phylogenies and the alignment, masking, model prediction and phylogeny generation steps were repeated.

6.3.2.3 STRUCTURAL PREDICTION AND FUNCTIONAL ANALYSIS

The structure of Pds1 was predicted from the *P. tetaurelia* protein sequence (PTETP600032001) using RaptorX (Källberg et al., 2012). This prediction used default settings and was based on a weighted combination of physico-chemical features of the amino acids sequence, 3D structural alignments, entropic modelling and domain prediction (Källberg et al., 2012). The predicted structure was then plotted from the PDB file using PyMOL (Delano, 2002).

Using the PDB structure from RaptorX, functional prediction was attempted using ProFunc (Laskowski et al., 2005), CombFunc (Wass et al., 2012) and PredictProtein (Rost et al., 2004) all with default settings.

6.3.3 dsRNA CROSS-TALK ANALYSIS - “eDicer”

In order to investigate the prevalence of “cross-talk” between host and endosymbiont a tool to analyse short sequence collisions between two sets of transcripts was created. “eDicer” is built around the Jellyfish K-mer counter (Marçais and Kingsford, 2011) and the K-mer Analysis Toolkit (KAT) (Clavijo et al., 2015). Using efficient K-mer hashing it allows the identification of shared K-mers between two sets of sequences. As Dcr1 in *P. tetaurelia* generates 23nt fragments, by identifying the number of shared 23-mers between two datasets e.g. the set of host and endosymbiont transcripts we can identify the number of potential RNAi “collisions” or cross-talks between the two species.

For each comparison between a set of query transcripts and a reference set of transcripts (endosymbiont transcripts in this case) the following values were calculated and tabulated:

- Number of K-mers in the query, in other words the length of the query in K-mers ($\frac{\sum_{n=1}^s \text{len}(x_n)}{w}$ for a set containing s transcripts and a K-mer size of w).
- Number of unique K-mers in the query, the non-redundant length of the query.
- Number of shared K-mers (“collisions”) between query and subject bin.
- Number of unique shared K-mers between query and subject bin.
- Shared K-mers normalised by subject length in K-mers.
- Shared K-mers normalised by the subject length in unique K-mers.
- Shared unique K-mers normalised by subject length in K-mers.
- Shared unique K-mers normalised by the subject length in unique K-mers.

In order to contextualise the number of collisions between the 2 host and 2 endosymbiont transcript sets I analysed the collisions between the two sets of endosymbiont transcripts with several other datasets.

These datasets were composed of predicted or sequenced transcriptomes from the following groups: bacteria, archaea, eukaryotes, green algae, and the ciliates. Sequences were selected to sample the breadth of the sequenced diversity of each group as fully as possible.

Specifically, 3 ciliate transcript sets were used *Paramecium tetaurelia*, *Tetrahymena thermophila* and *Oxytricha trifallax* along with 5 green algae *Chlamydomonas reinhardtii*, *Cocomyxa subellipsoidea* C-169, *Chlorella variabilis* NC64A, *Micromonas pusilla* RCC299, and *Ostreococcus lucimarinus*. The eukaryote dataset was composed of 58 transcript sets (??), the bacterial 130 (??), and the archaea 89 (??).

Tabulated values were then analysed statistically using standard python tools outlined in the methods chapter.

In the course of this work improvements made to KAT were submitted and merged back into the core KAT development codebase (<https://github.com/TGAC/KAT>)

6.4 RESULTS

6.4.1 INDUCTION OF RNAi

There was a total failure to induce RNAi related phenotypes in feeding experiments for any of the vectors. Despite observation for several days, continuous refeeding with transformed *E. coli* (and RT-PCR proof of dsRNA expression) none of the *P. bursaria* CCAP 1660/12 cultures displayed any altered phenotypes as a consequence of feeding experiments.

Microinjection was attempted using just the PGM and epi2 constructs due to time constraints. *P. bursaria* tended to burst after a single injection attempt. As a control, GFP was also attempted to be injected into the MAC.

Unfortunately, despite numerous attempts it was never possible to observe fluorescence localised to the MAC. This suggests microinjection was never successfully achieved.

6.4.2 RNAI PATHWAY COMPONENTS

6.4.2.1 DCR1

A single Dcr1 orthologue was clearly identified in each of the two *P. bursaria* host transcriptomes. Phylogenetic analysis (fig. 6.4.1) recovered a phylogeny matching the established ciliate taxonomy ((Aury et al., 2006; Fokin et al., 2004; Swart et al., 2013)) with strong support for *P. bursaria* as the the outgroup to the rest of the *Paramecium* clade.

6.4.2.2 PDS1

There were no hits for Pds1 in any of the 3 *P. bursaria* datasets (both transcriptomes and the genome). There were homologues in each of the other *Paramecium* species i.e. *P. sexaurelia*, *P. biaurelia*, *P. primaurelia*, *P. multimicronucleatum* and *P. caudatum* but not *T. thermophila* or *O. trifallax* (fig. 6.4.2).

As this protein has no assigned function based on sequence homology (Marker et al., 2014; Carradec et al., 2015) but potentially plays an important role in the uptake of RNA from vacuoles at attempt was made to infer function by structural analysis.

The structure of Pds1 (fig. 6.4.3) was predicted from the *P. tetraurelia* sequence via RaptorX (Källberg et al., 2012). Unfortunately, no function could be assigned to this structure using ProFunc (Laskowski et al., 2005), CombFunc (Wass et al., 2012) or PredictProtein (Rost et al., 2004). No structural hits were found against known enzyme active sites, ligand-binding sites or DNA-binding templates in ProcFun. Therefore, no function could be assigned to this protein.

6.4.2.3 CID

Two Cid orthologues were identified in each *P. bursaria* transcriptome. Additionally, a closely related Cid3 orthologue was identified in the other *Paramecium* sequences.

Phylogenetic analysis (fig. 6.4.4) showed an unclear picture with both the Yad1g Cid peptides and 1 of the SW1 peptides branching with moderate support (86.7% bootstraps and a PP of 0.69) within a clade composed of Cid1 and Cid3. Unfortunately, there was poor support (59.1%/0.54) for the *P. bursaria* sequences forming a sister to these clades therefore, their exact placement is unclear. Additionally, the other SW1 orthologue branched as the outgroup to all remaining Cid with high support (100%/1.00).

This suggests that the orthologues present in Yad1g and one of the orthologues in SW1 may be the unduplicated ancestor to Cid1 and Cid3 (named Cid1-3 for convenience).

In general this phylogeny is consistent with a scenario in which a single ancestral Cid has undergone duplication resulting in Cid2 and a Cid1-3 ancestor either before or after the branching of *P. bursaria*. If this divergence

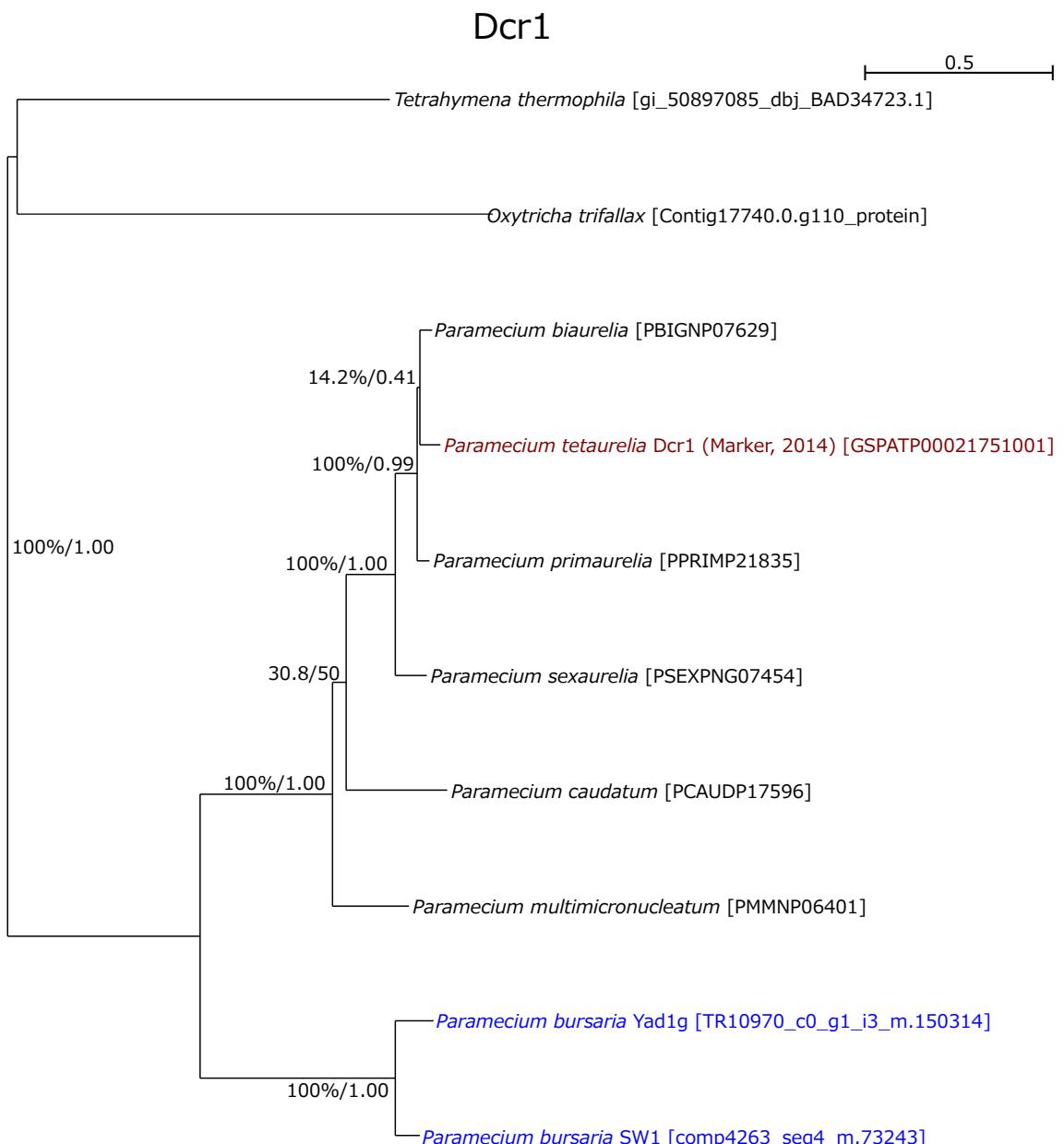


Figure 6.4.1: Dcr1 Phylogeny (1196 sites) inferred using RAxML with LG+G+F and 1000 non-rapid bootstraps. Bayesian PP were inferred using MrBayes with 2 runs of 4 chains run for 2,000,000 generations (5% burnin) and the LG+G model. *P. bursaria* peptides are highlighted in blue whereas *P. tetaurelia* components identified by (Marker et al., 2014) are indicated in red. Phylogeny is largely consistent with established ciliate phylogenies.

Psd1

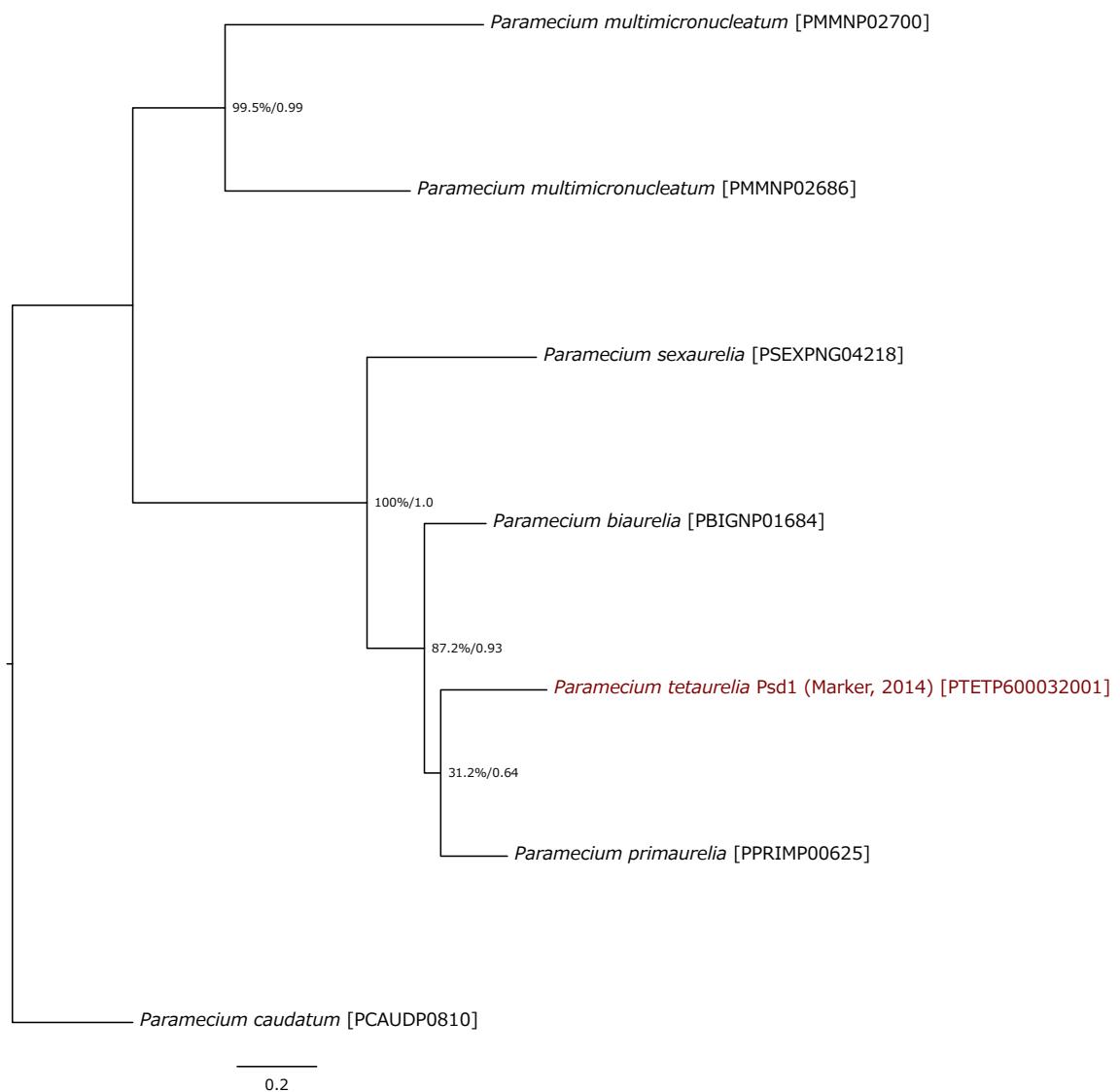


Figure 6.4.2: Pds1 Phylogeny (424 sites) inferred using RAxML with VT+G+F and 1000 non-rapid bootstraps. Bayesian PP were inferred with MrBayes (2 runs of 4 chains for 2,000,000 generations, 5% burn-in) with VT+G and annotated. This phylogeny is consistent with the general taxonomy of the ciliate clade.



Figure 6.4.3: RaptorX Predicted Structure of *P. tetaurelia* Psd1 protein (PTETP600032001). No functional annotations could be made using this structure.

occurred before this speciation then Cid₂ has been lost in *P. bursaria*. Regardless, there is a clear subsequent duplication and divergence of the Cid₁₋₃ ancestor into the modern Cid₁ and Cid₃ between the *P. bursaria* and *P. caudatum* branches.

Interestingly, the timing location of these events and the presence of all 3 Cid homologues in *P. caudatum* (which shares the single ancient WGD with *P. bursaria*) suggests that this pattern is not directly related to the polarised position of WGD in *Paramecium*.

6.4.2.4 RDR

Two putative sequences were identified by Rdr₁ and Rdr₂ searches in each *P. bursaria*. Phylogenies of these sequences (fig. 6.4.5) indicate that *P. bursaria* has an orthologue of Rdr₂ (strong support (99.9%/1.00) of an outgroup to the Rdr₂ sequences). The other sequences from both transcriptomes branch basally to the other *Paramecium* RdRPs with moderate/weak support and may nor may not be an orthologue of Rdr₁.

Rdr₃ analysis didn't identify any hits in the *T. thermophila* or *O. trifallax* outgroups but did find an orthologue in both *P. bursaria* transcript sets. These branched as a sister to the other *Paramecium* with strong support suggesting that they may be orthologous to the *P. tetraurelia* Rdr₃. The phylogeny recapitulated the taxonomy of the ciliates (fig. 6.4.6). However, the lack of homology to the other Rdrs indicates that this Rdr may have been an independent innovation arising basally to the *Paramecium* clade and is unrelated to Rdr₁ and Rdr₂.

6.4.2.5 PIWI

There were a large number of Piwi detected homologues across the datasets. Specifically, 16 Piwi in the *P. bursaria* SW-1 transcriptome and 5 in the partial genome, and 17 in the *P. bursaria* Yad1g transcriptome. Due to the large size of this family, large number of paralogues and relatively short sequences phylogenetic inference of these sequences proved largely intractable. There are Piwi homologues present in *P. bursaria* although their exact relation and function is unknown.

6.4.3 dsRNA COLLISIONS

The total number of unique collisions between the endosymbionts and the various classes of subject transcriptomes were plotted (fig. 6.4.7). This showed next to no collisions with Archaea, moderate levels of total collisions with Bacteria and generally higher levels of collision with the Eukaryotes. The elevated number of collisions with eukaryotic transcriptomes is not surprising given their phylogenetic position and gene content. However, the difference in total collisions between archaea and bacteria is potentially interesting. It is possible this reflects the sequencing bias in archaea towards extremophiles which often have compositional adaptations.

Additionally, as might be expected from their close relationship with the green algal endosymbionts some of the most frequent collisions were with the other green algae. Interestingly, the 3 ciliate species displayed a generally low number of collisions but there was moderate to high levels of collision against the two *P. bursaria*

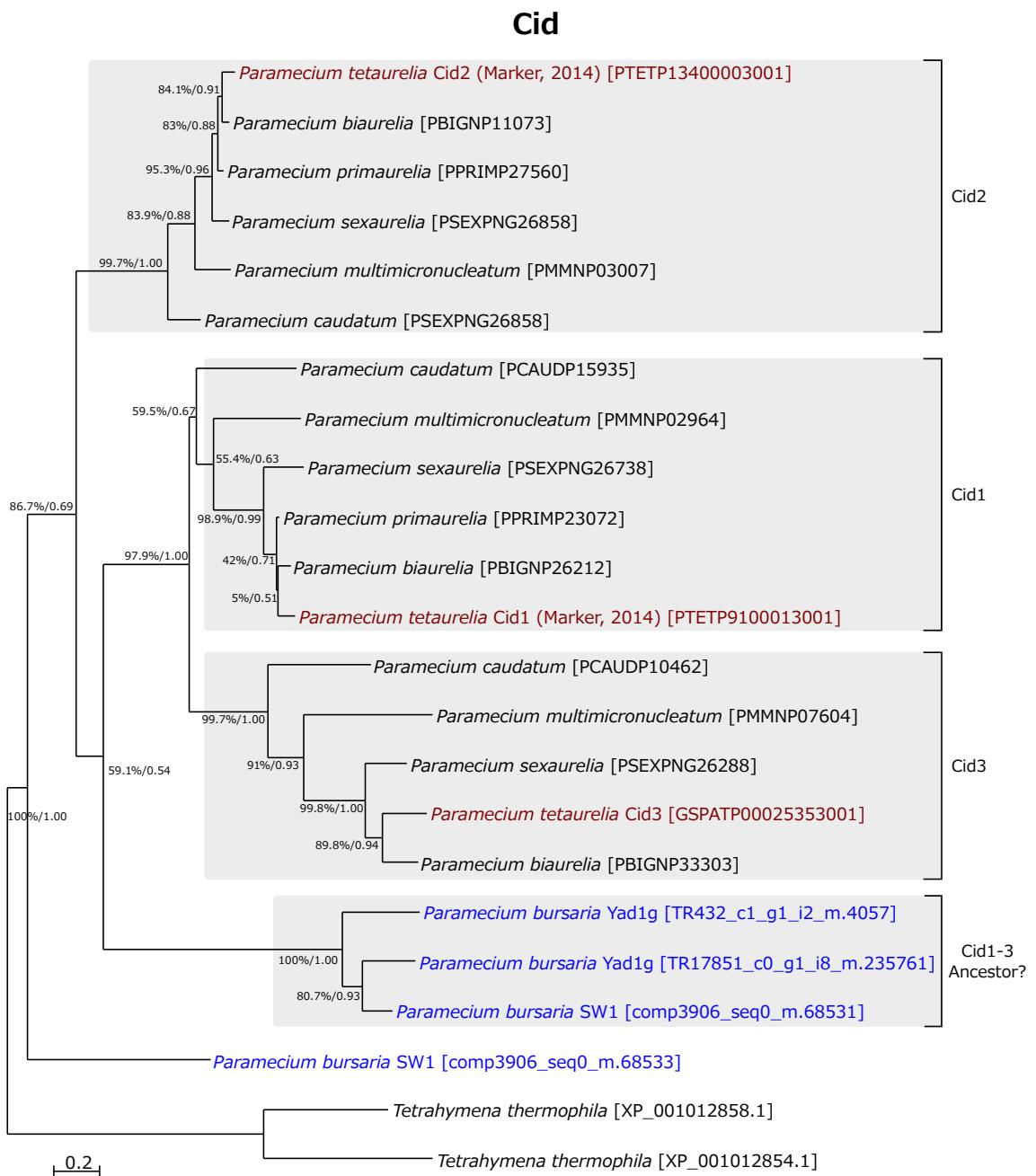


Figure 6.4.4: Phylogeny of Cid1, Cid2, and Cid3 (268 sites) inferred using RAxML with rtREV+G and 1000 non-rapid bootstraps. Bayesian PP were inferred using MrBayes (2,500,000 generations with 2 runs of 4 chains and a 10% burn-in). Phylogeny shows a potential orthologue of an ancestral pre-divergence version of Cid1 and Cid3 (named Cid1-3) in *P. bursaria* Yad1g and SW1 and an uncertain Cid orthologue possibly related to Cid2 depending on timing of the Cid1-3 and Cid2 divergence in *P. bursaria* SW1.

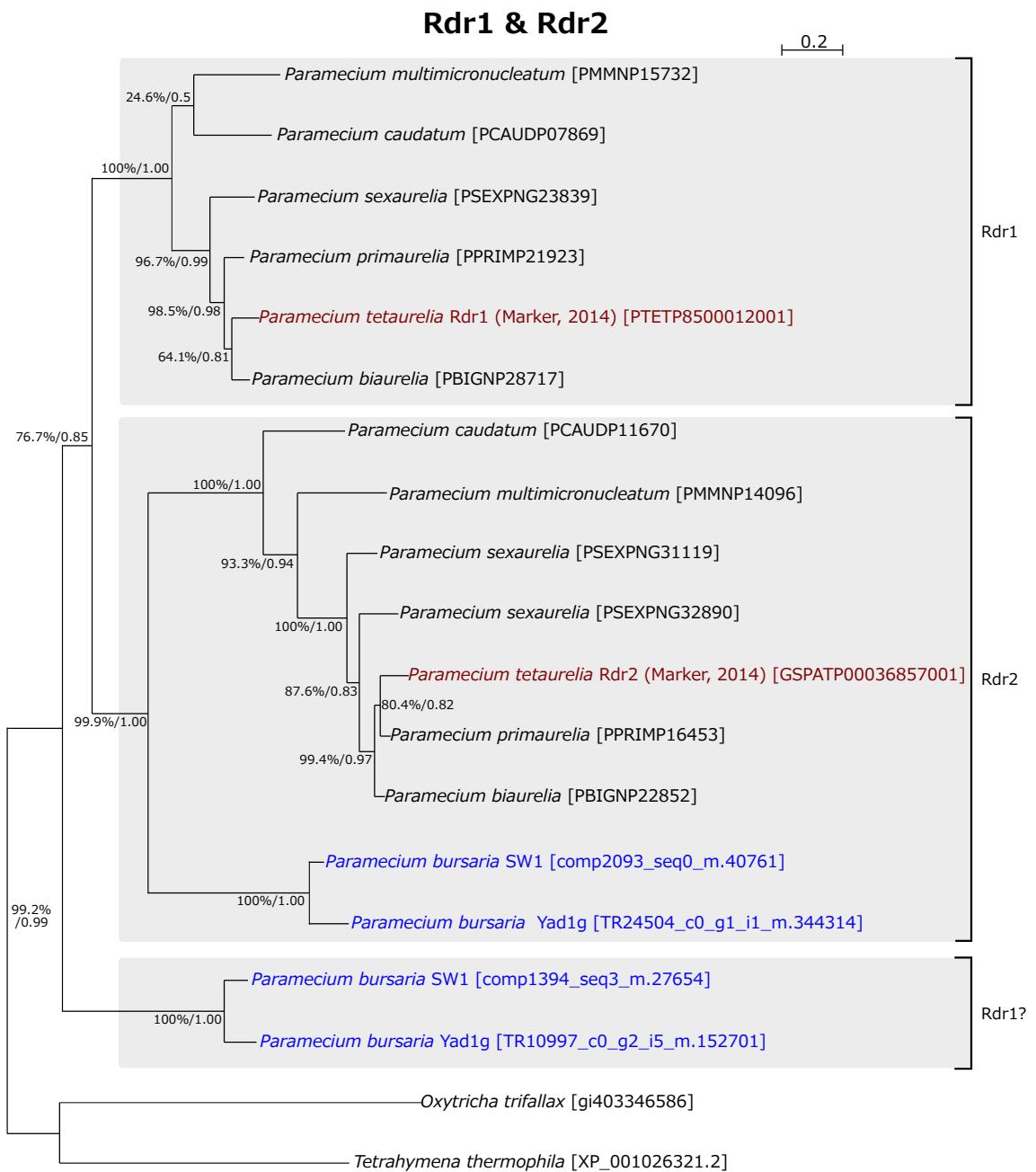


Figure 6.4.5: Phylogeny of Rdr1 and Rdr2 (691 sites) inferred using RAxML with LG+G+F and 1000 non-rapid bootstraps. Bayesian PP were inferred using MrBayes (2,000,000 generations with 2 runs of 4 chains and a 5% burn-in) and annotated onto the RAxML phylogeny. Phylogeny shows a homologue of Rdr2 in *P. bursaria* as well as a potential ancestral or Rdr1 homologue.

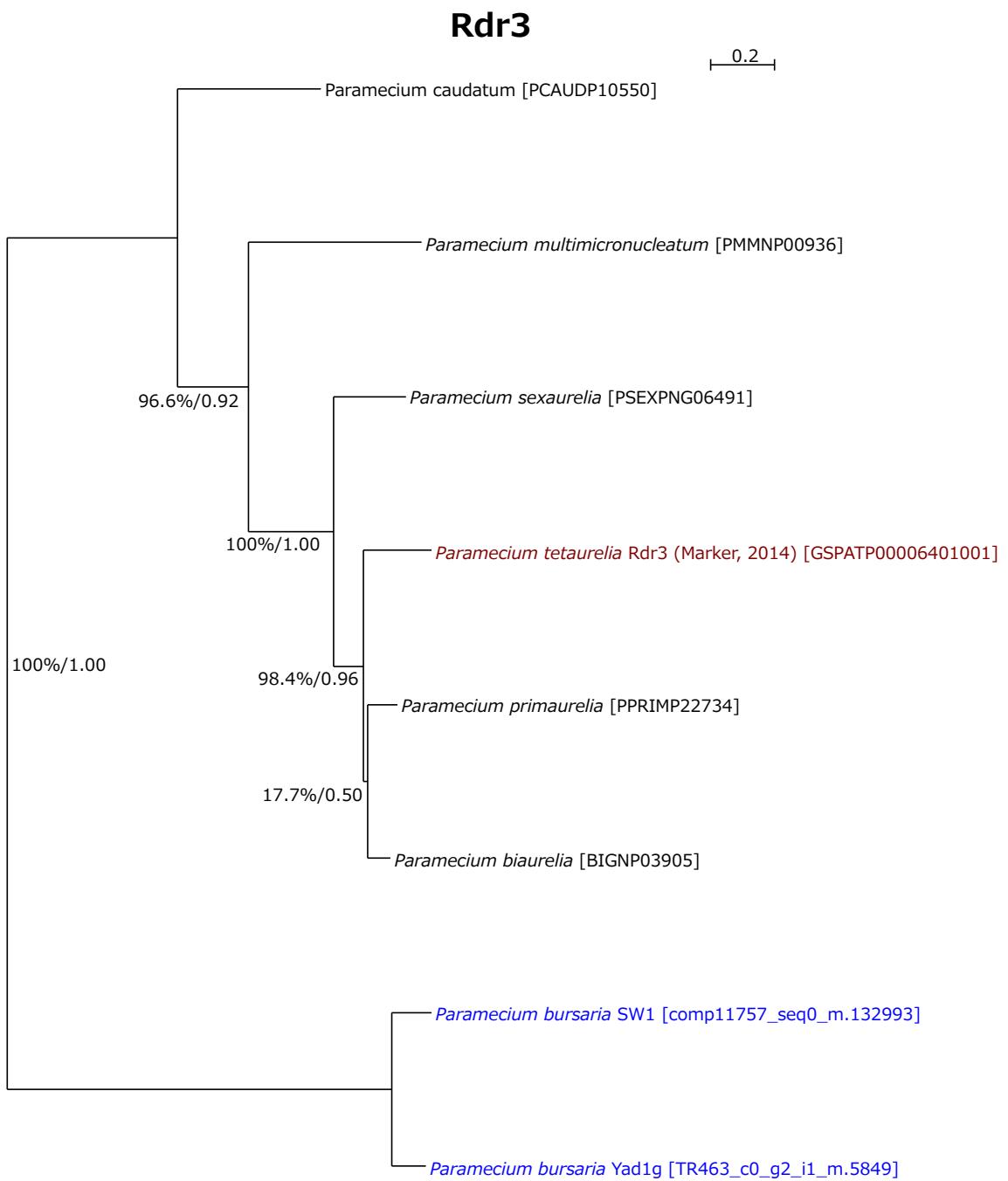


Figure 6.4.6: Phylogeny of Rdr3 (432 sites) inferred using RAxML with JTT+G+F and 1000 non-rapid bootstraps. Bayesian PP were inferred using MrBayes (2,000,000 generations with 2 runs of 4 chains and a 5% burn-in) and annotated onto the RAxML phylogeny. Phylogeny shows the presence of a likely Rdr3 in *P. bursaria*.

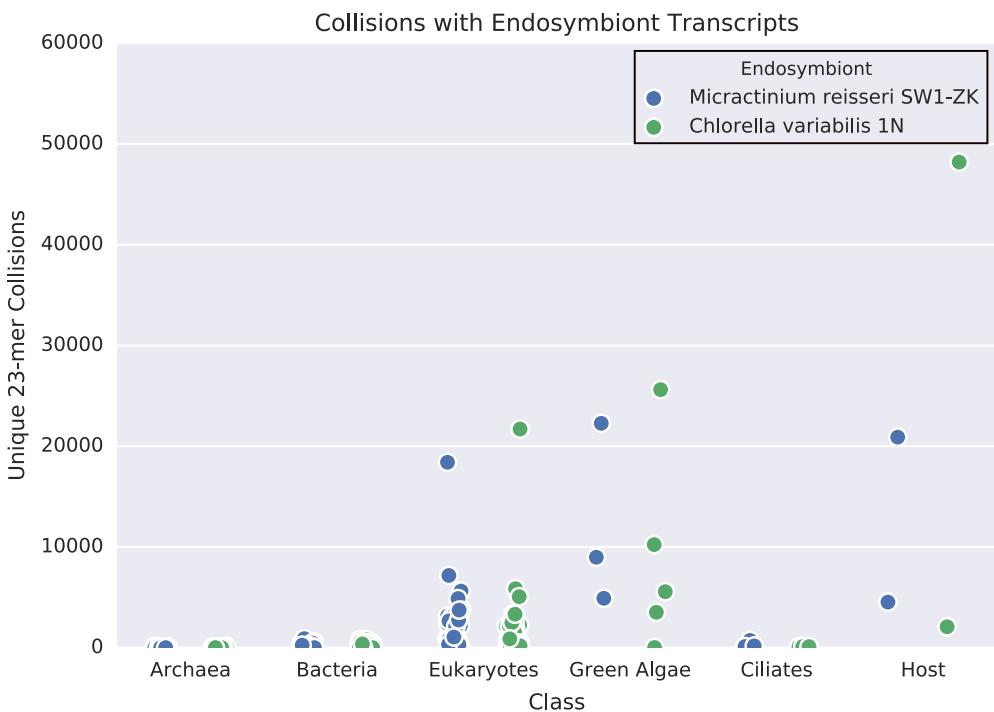


Figure 6.4.7: The number of unique 23-mer collisions between transcripts from the *C. variabilis* 1N and *M. reisseri* SW1-ZK endosymbionts and different classes of subject transcripts. The y-axis is truncated to better separate the classes however, the the only taxa that had more than 6000 unique collisions were *Coccomyxa variabilis* NC64A and *Chlorella subellipsoidea* C-169 , *Arabidopsis thaliana*, *Chlamydomonas reinhardtii* and transcripts from the two host bins *P. bursaria* SW1 and *P. bursaria* Yad1g. This suggests RNAi-cross talk could be occurring between the host and endosymbiont and would likely occur at higher rates than occurs between *Paramecium* and their bacterial prey.

host transcriptomes. This potentially indicates a higher level of collision between the active host genome rather than its total genome.

The collision level was relatively consistent for both endosymbionts when compared to the same subject (fig. 6.4.8). However, there are visible exceptions where a given subject transcriptome has far more collisions against one of two endosymbiont transcripts than the other. This can be seen in the line in the pairplots with sharp angles instead of being close to level.

This is particularly obvious in the comparison of collisions between the host endosymbiont pairs (fig. 6.4.9). There is a considerable difference in collisions for the *C. variabilis* 1N endosymbiont from the Yad1g1N culture with next to no collisions against the *P. bursaria* from the other culture but a high level of collisions against its own host. Interestingly, *M. reisseri* was relatively consistent across both hosts with only slightly more hits to its own host. This suggests a potential problem in the binning of endosymbiont and host transcripts, especially in the *P. bursaria*-*C. variabilis* Yad1g1N transcriptome.

In order to test to what degree the number of collisions was related to the length of the subject query, a simple linear regression was conducted (fig. 6.4.10). This demonstrates there is at least a partial linear dependence between the length of the query and the number of collisions as would be expected.

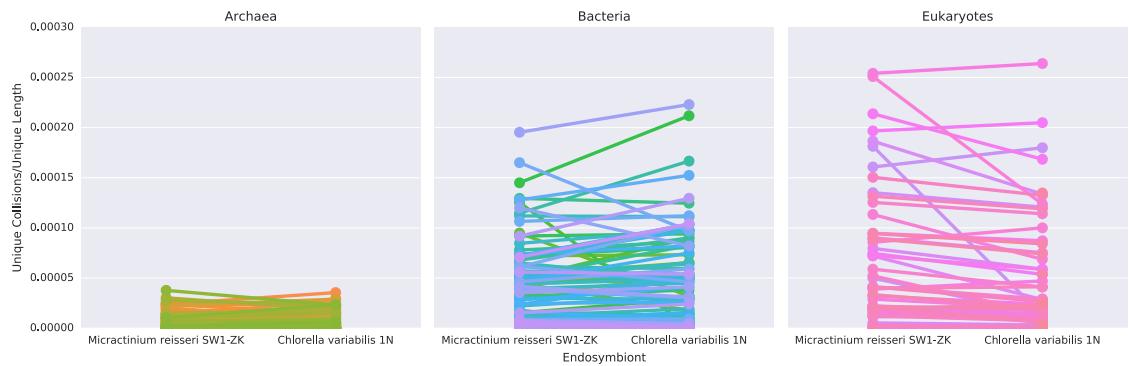


Figure 6.4.8: Pair-plot of the normalised collisions in Archaea, Bacteria and Eukaryotes showing the relative consistency of the number of hits between the two endosymbionts. Lines join the number of collisions to the same subject transcriptome in the different endosymbiont transcriptomes. Colour is merely for illustration purposes and has no significant meaning. The plot shows that overall the collisions are relatively consistent but there are aberrations where one endosymbiont has far more collisions to a given subject than the other.

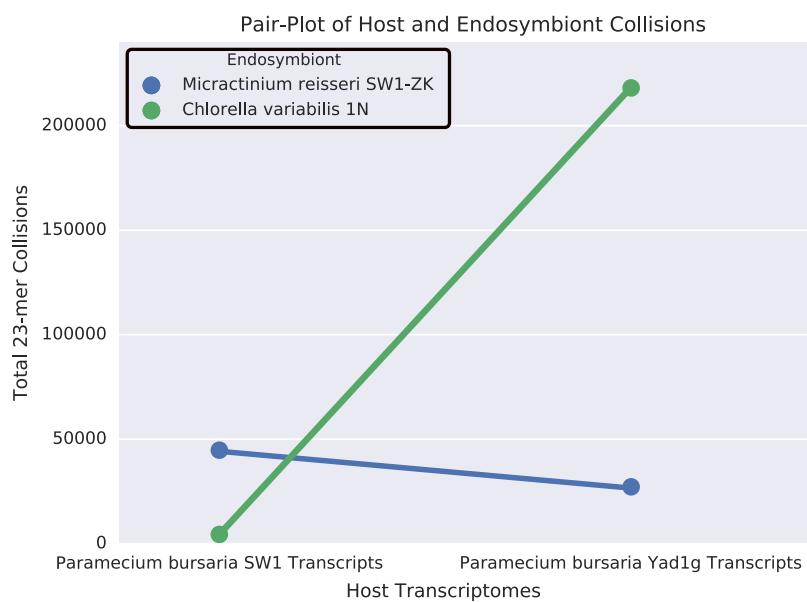


Figure 6.4.9: Pair-plot of the total number of collisions between the two endosymbiont transcriptome sets and the two host transcriptome sets. *M. reisseri* has a relatively consistent number of collisions against both hosts, however, the *C. variabilis* transcriptome has a significantly higher number of collisions against its own endosymbiont. This suggests potential issues with the binning of transcripts in the *Yad1g1N* transcriptome.

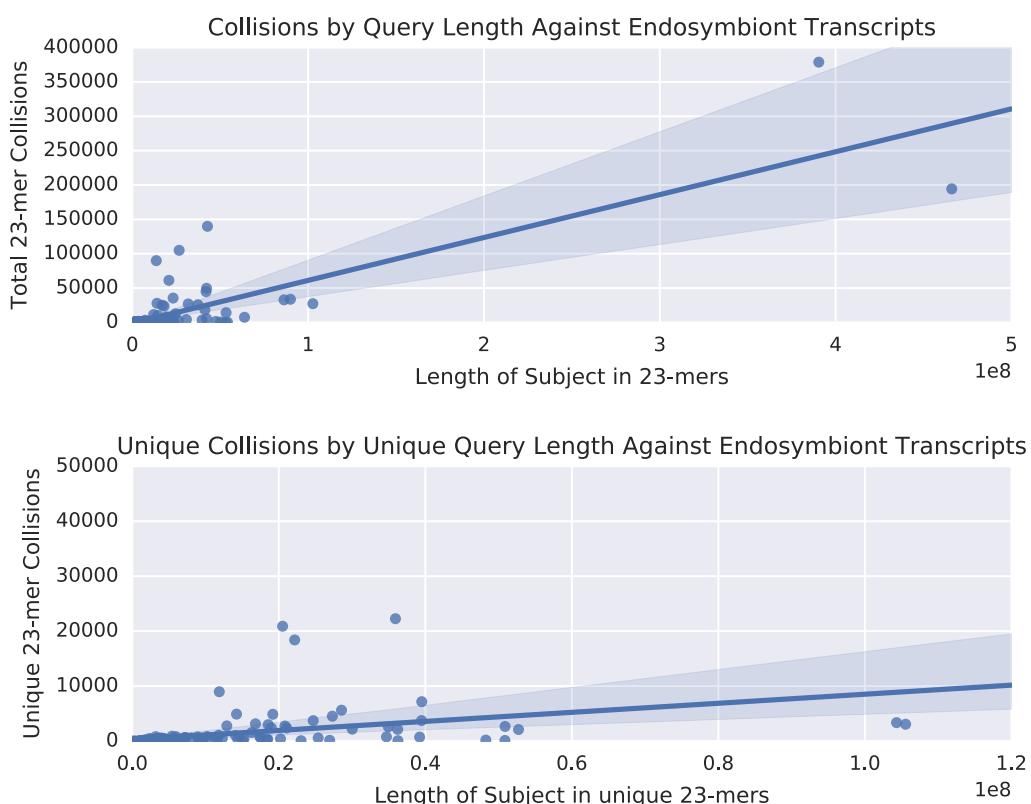


Figure 6.4.10: Linear Regressions of the relationship between the number of collisions as query length increases. Features points from collisions with both *C. variabilis* 1N and *M. reissier* SW1-ZK. Dark blue cloud indicates 95% confidence intervals. The top plot shows total collisions against total length whereas the bottom analysis only looks at unique collisions against unique length so theoretically ameliorates the effect of repetitive transcripts/many isoforms. Both show a clear if noisy (correlation coefficients of 0.18085 and 0.02811 respectively) linear relationship and thus support that the size of a transcriptome plays a role in the number of collisions.

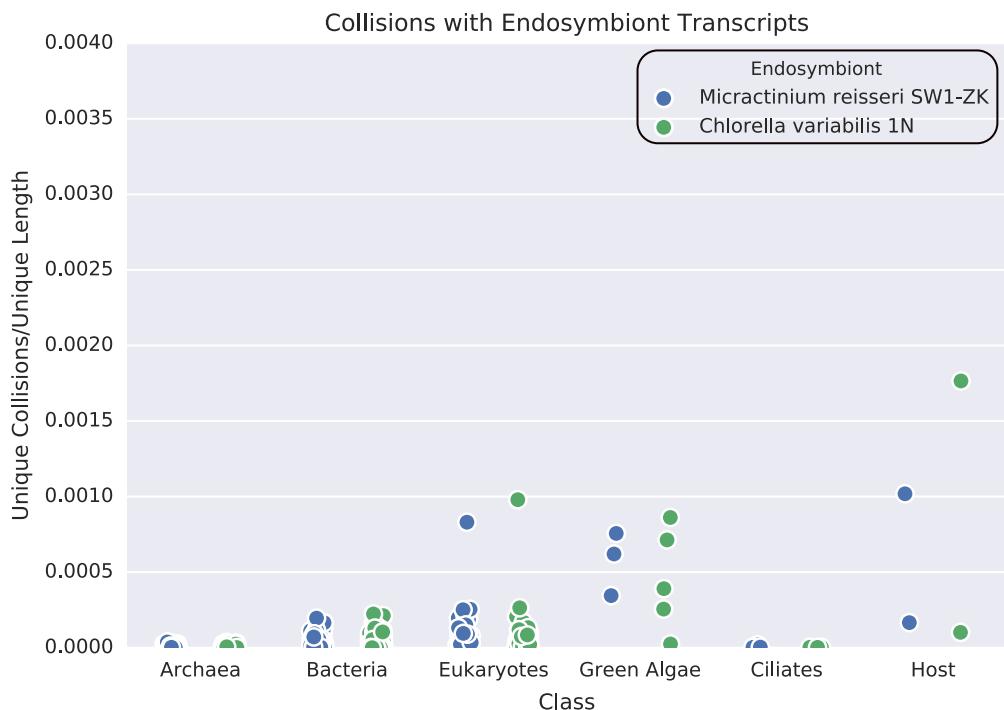


Figure 6.4.11: The number of collisions between transcripts from the *C. variabilis* 1N and *M. reisseri* SW1-ZK endosymbionts and different classes of subject transcripts normalised by the unique length of those subjects.

Normalising the unique number of collisions by the unique length of the subject predicted transcriptome led to some interesting results (fig. 6.4.11) Collisions with Eukaryotes largely were reduced to being on-par with collisions against Bacteria. Despite collisions against the ciliates relatively disappearing, there was still considerable levels of potential RNAi cross-talk against the host transcriptomes and green algae.

6.5 DISCUSSION

6.5.1 NO dsRNA RNAI INDUCIBLE PHENOTYPES IN *P. BURSARIA* SW1

Despite numerous attempts, all feeding experiments failed to induce any of the RNAi knockout phenotypes. RT-PCR tests (not shown) with the Bug22 and BBS7 constructs demonstrated that the *E. coli* was successfully transformed and could inducibly express the dsRNA. This indicates that dsRNA is either incapable of escaping the digestive vacuole or that *P. bursaria* does not have an active pathway for exogenous dsRNA induced RNAi.

The potential failure of microinjection of dsRNA directly into the MAC to elicit RNAi phenotypes would support the absence of an active dsRNA induced RNAi pathway. However, the high methodological difficulty involved in identifying and injecting the MAC without lysing the cell (fig. 6.5.1) means microinjection may have only failed to induce RNAi due to failure to correctly microinject *P. bursaria*. *P. bursaria* was particularly prone to lysis relative to *P. tetraurelia* and there were greater issues with trichocysts blocking the microinjector. Attempts made to inject GFP into the MAC also failed to generate fluorescence localised to the MAC. This indicates that

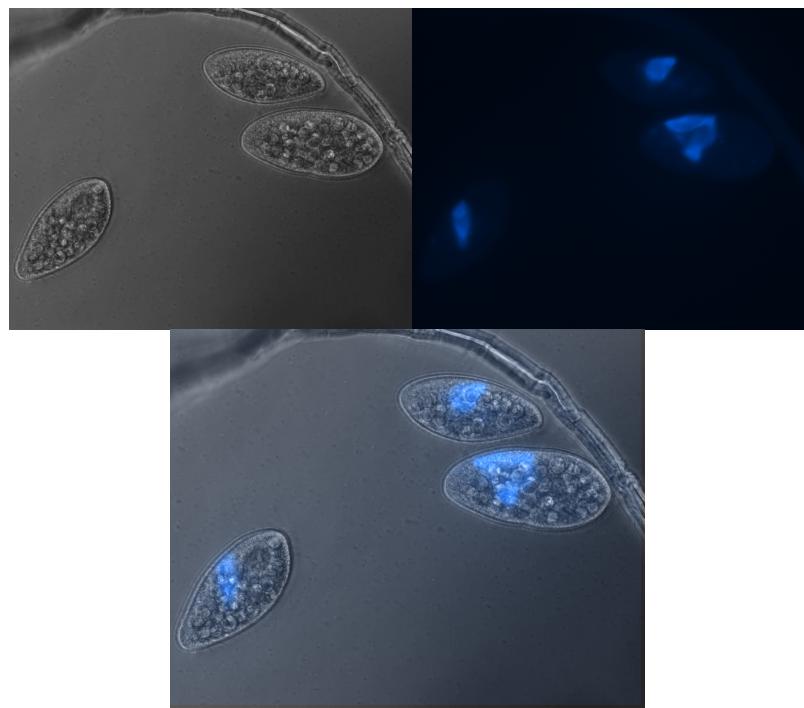


Figure 6.5.1: DAPI stained *P. bursaria* MAC to demonstrating how difficult it is to accurately identify the limits of the MAC for microinjection. Top left panel shows 3 *P. bursaria* cells under light microscopy as they appear when attempting microinjection. Top right shows fluorescence from DAPI staining to demonstrate where the MAC nuclei are location. Bottom panel is an artificial overlay of the light microscopy and DAPI staining.

the failure of microinjection is primarily technical and cannot be used to make inferences about the state of the RNAi pathway in host.

Unforunately, the alternative transgene RNAi methodology ([Galvani et al., 2001](#)) (which was not attempted) also involves microinjection of the MAC with the transgene construct itself. Therefore, even if the transgene pathway is present and active it may still not be possible to reliably induce RNAi in *P. bursaria*.

6.5.2 MISSING COMPONENTS OF RNAI PATHWAY IN *P. BURSARIA*

An investigation into the RNAi pathway components identified by ([Marker et al., 2014](#)) revealed the absence of one required component for the exogenous dsRNA pathway (*Pds1*) and depending on the function of a putative ancestral *Cid* protein the absence of a factor in the common pathway or another exogenous dsRNA pathway (fig. 6.5.2).

Pds1 is totally absent outside of the post-*P. bursaria* *Paramecium* clade. A phylogenetic analysis of *Pds1* sequences (fig. 6.4.2) recapitulated the established *Paramecium* taxonomy (fig. 1.2.2). This suggests that *Pds1* was either acquired after the divergence of *P. bursaria* and *P. caudatum* or was lost in *P. bursaria*. The pattern of paralogues would more likely support the former scenario. The lack of paralogues (with the exception of a terminally duplicated *P. multimicronucleatum* copy) in the *P. aurelia* complex species is interesting. As the presence of *Pds1* in *P. caudatum* indicates that this should have undergone duplication during the 2 subsequent WGD events. Potentially, this represents serial losses in these species.

It is possible that *Pds1* is present and just hasn't been recovered in the partial transcriptomes because it is not

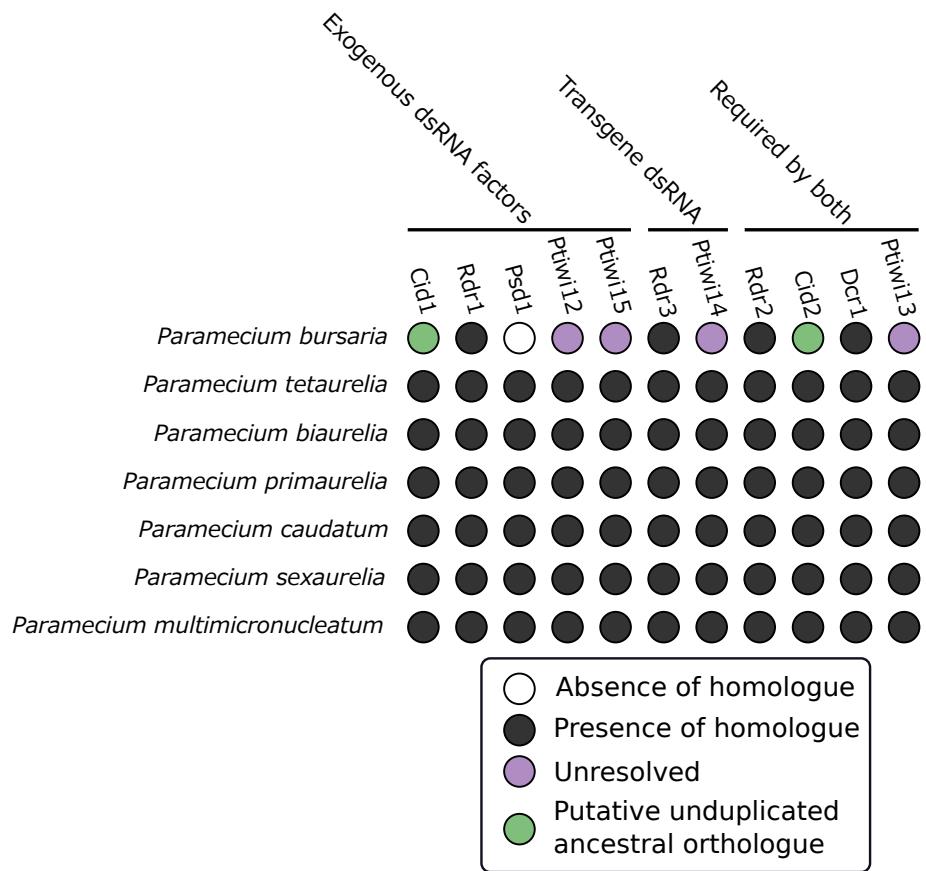


Figure 6.5.2: Coulson plot showing the absence/presence of RNAi pathway factors identified in *P. tetaurelia* (Marker et al., 2014) across the *Paramecium* clade.

being transcribed (or is transcribed at a very low level) during endosymbiosis. It is also possible it is missing in the partial genome due to the incompleteness of this data. However, the combination of being missing in all 3 datasets as well as any non-*Paramecium* ciliates indicates that it is likely not present in *P. bursaria*.

The Cid proteins are difficult to resolve, either *P. bursaria* has an undiverged ancestral version of the Cid proteins or has a Cid1-3 ancestor and has secondarily lost Cid2. This depends on the timing of the Cid2 and Cid1-3 divergence, if it occurred after the branching of *P. bursaria* from the rest of the clade then the latter scenario is more likely and vice versa.

If the latter scenario is true then Cid2 has been lost and if the *P. bursaria* RNAi pathways require the same components as *P. tetaurelia* this may mean both the transgene and dsRNA pathways might not be active. Additionally, if the Cid1-3 ancestor or the ancestral undiverged Cid does not have function of *P. tetaurelia*'s Cid1 then the exogenous dsRNA pathway may not be active in *P. bursaria*.

The unresolved Piwis also represent potential issues with the RNAi pathway but the only way to thoroughly investigate the roles of these proteins would be targeted mutagenic screening or a similar approach.

The low levels of sequence homology suggests independent innovation of Rdr3 and potentially sheds doubt on its relationship to the ancestral Rdr that this analysis confirms was likely present in was present in the first WGD (Marker et al., 2014).

Interestingly, the general disposition of paralogues across the *Paramecium* clade does not recapitulate the 3

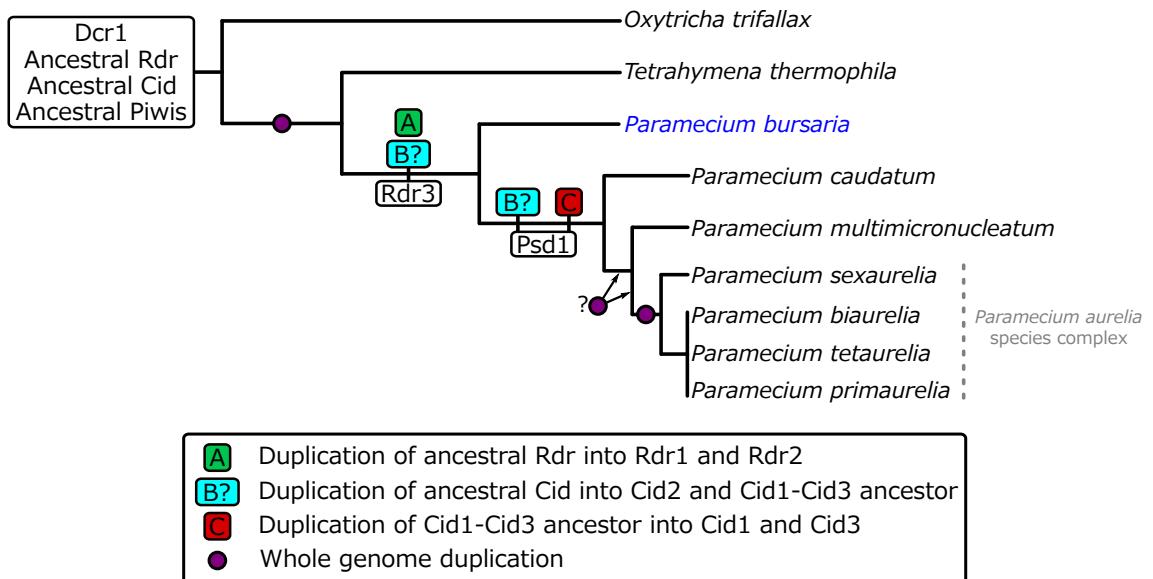


Figure 6.5.3: Diagram showing the *Paramecium* clade with *Tetrahymena* outgroup showing the putative evolutionary scenarios behind the currently observed distributed of RNAi factors.

established WGD events in this group. For example, there is a relatively consistent number of paralogues of all components, especially RdRPs and Cid proteins in *P. caudatum* despite this clade having undergone the same number of WGD as *P. bursaria* and not sharing the 2 recent WGD with the majority of the remaining *Paramecium*

6.5.3 ENDOSYMBIONT “COLLISION” HYPOTHESIS

Hypothetically *Paramecium bursaria* may have deactivated/lost feeding induced RNAi (specifically the uptake of RNA from vacuoles) as a consequence of the greater levels of potentially deleterious cross-talk between it and its eukaryotic green algal endosymbionts. As an exogenous RNAi response is not essential for viability in *P. tetaurelia* (Marker et al., 2014) loss of this system in *P. bursaria* may not have a high fitness cost.

On first glance the high levels of collisions between endosymbiont transcripts against host and eukaryote classes in general support this hypothesis (fig. 6.4.7). However, there were low levels of collisions between endosymbiont transcripts and ciliate transcripts indicating that the collision levels observed between host and endosymbiont require an additional explanation. The first scenario is that the active host transcriptome during endosymbiosis is not representative of all possible host transcripts and features a much higher level of collision with the endosymbiont. The second scenario is that the levels of collisions between host and endosymbiont actually reflects misbinning of endosymbiont transcripts as belonging to the host. A comparison of the collisions of each endosymbiont against its own host and the other host *P. bursaria* (fig. 6.4.9) suggests that binning error may explain some of this difference and that the binning the Yad1g1N transcriptome has potentially more issues than the SW1-ZK (CCAP 166o/12) transcriptome. This might be explained by the fact that only half the sequenced libraries in Yad1g1N actually contained the *C. variabilis* 1N endosymbiont as this dataset originated from an analysis of transcriptomic profiling of *P. bursaria* with and without its endosymbiont (Kodama et al., 2014).

There is a linear relationship between the unique size of a transcriptome in terms of 23-mers and the number of collisions (fig. 6.4.10). This is expected as a longer set has a higher probability of a random match by chance.

When the number of collisions is normalised by the length of the subject (fig. 6.4.11) the level of collisions from eukaryotes is largely on-par with those of bacteria. However, as eukaryotes have larger genomes the un-normalised number of collisions is more reflective of biological reality. Indeed, the size and diversity of their transcriptome may be the reason why eukaryotic cross-talk is potentially more problematic than that of bacterial cross-talk.

It might be interesting to test the linear relationship between phylogenetic relatedness and the number of K-mer collisions. For example, phylogenetic distances between the endosymbiont and various taxa could be derived from an established published multi-gene analysis that includes these species and regression conducted using this as a feature.

One down-side of the efficient K-mer hashing based “eDicer” design is that it only finds exact matches. RNAi has been found to not always require exact sequence matches to induce knock-down ([Elbashir et al., 2001](#)) therefore, this analysis potentially misses a large amount of near-identical collisions. Fortunately, it is probably a relatively safe assumption that the number of identical collisions correlates strongly with the number of near-identical collisions.

Finally, while this may not have thoroughly resolved the question of matches eDicer may form a useful tool in the rapid screening of off-target effects in the RNAi analyses in different organisms. It is considerably more efficient than cutting and alignment based methods such as that offered on ParameciumDB ([Arnaiz and Sperling, 2011a](#)).

6.6 CONCLUSIONS

RNAi induced phenotypes could not be created in *P. bursaria* SW1 from CCAP1660/12 via feeding experiments. The absence of Pds1 in *P. bursaria* offers a potential explanation for this as this protein has been implicated in playing some undefined role in the uptake of RNA (dsRNA or ssRNA) from digestive vacuoles ([Carradec et al., 2015](#)). As “eDicer” identified that there are a large number of 23-mer collisions between *P. bursaria* transcripts and eukaryotic transcriptomes (especially the green algal endosymbionts) the loss/deactivation of the uptake of RNA from digestive (or potentially perialgal) vacuoles could be a consequence of a eukaryotic endosymbiont in *P. bursaria*. The relatively lower levels of RNAi cross-talk between *Paramecium* and bacterial endosymbionts and/or food species may prove less deleterious than eukaryotic cross-talk.

Due to Psd1 only likely being involved in uptake from a food vacuole it is possible that dsRNA could still be induced by direct microinjection. Similarly, microinjection of transgenes may still be possible. Unfortunately, microinjection has proven difficult technically in *P. bursaria*. Further optimisation of the experimental method and training is required to thoroughly test the activation or deactivation of injected dsRNA or transgenes.

Alternatively, the potential ancestry of the Cid proteins in *P. bursaria* may indicate a deactivation/absence of RNAi by the pathways identified by ([Marker et al., 2014](#)) in *P. bursaria* depending on the relative functionality of this ancestral form. Further analysis of RNAi systems in *P. bursaria* and other *Paramecium* species would be required to further answer this question.

"a typical symbiotic Chlorella strain common to all P. bursaria strains does not exist"

- Reisser et al. (1988)

7

Conclusions and Recommendations

There were 3 main objectives to this research:

- Assessing the utility of *P. bursaria* and its endosymbioses with green algae as a model organism for the study of the evolution of endosymbiosis.
- Generation of “omic” resources to inform further analysis of this system
- Investigating the the utility and feasibility of current MDA-based single cell genomic and transcriptomic sequencing technique in the analysis of complex multimember single-celled eukaryotic systems.

Review of literature established that *Paramecium bursaria* and its 4 algal endosymbionts: *Chlorella variabilis*, *Chlorella vulgaris*, *Coccomyxa* sp., and *Micractinium reisseri* make theoretically good model organisms for the study of this process due to the existence of extensive. This is due to them being believed to share:

- a well-developed background literature.
- facultative endosymbioses allowing elimination of the endosymbiont and re-introduction experiments.
- the potential of a functional and tractable RNAi system for hypothesis testing via gene transcription knock down.
- easily culturable and diverse available cultures.

However, with the exception of one transcriptomic analysis (Kodama et al., 2014) there have been no analyses of these systems using contemporary transcriptomic, metabolomic or genomic techniques.

An investigation of elimination of the endosymbiont in the *P. bursaria* -*M. reisseri* culture revealed that this endosymbiosis may represent an obligate system, at least on the part of the *P. bursaria* SW1 host. Even though 3 elimination methods and a range of treatment concentrations were attempted (to minimise the risk that the host death was related a particular susceptibility to a certain treatment) all methods resulted in the same eventual host death. Future work is required to establish whether *M. reisseri* algae are also obligate endosymbionts. If this mutually obligate system is the case then metabolic co-dependence may have become fixed between *P. bursaria* and *M. reisseri*. This can be tested by testing whether an axenic culture of *M. reisseri* can be established. This could be achieved by exploiting robustness of the chitinous cell wall of the algae relative to the *Paramecium* membrane. Gentle agitation would allow the lysis of the endosymbiont without the lysis of a significant number of the endosymbionts. The difficulty in this would be optimising the culture conditions for *M. reisseri* as these endosymbiont algae are known to be relatively fastidious (Hoshina and Imamura, 2009).

As the Yad1g1N culture has previously been established as a facultative further work could use the Yad1g1N and CCAP1660/12 transcriptomes created here to attempt to identify key differences between them. Identifying these differences may pinpoint the mechanism by which metabolic co-dependence becomes fixed in *P. bursaria* - green algal endosymbioses.

The ITS2 sequence analysis of the CCAP 1660/12, and CCAP 1660/13 cultures revealed that these cultures likely contained the same *M. reisseri* endosymbiont and not a *Coccomyxa* endosymbiont as described in the culture collection. The identity of the Yad1g1N endosymbiont as *C. variabilis* 1N was also confirmed. A taxonomic analysis of the *Paramecium* species should also be conducted targeting the single MIC copy of the rDNA to confirm the host identities in the primary cultures.

This ITS2 analysis also established that the photobionts in the CCAP 1660/12, CCAP 1660/13 and NBDJ Yad1g1N cultures most likely form clonal photobiont cultures within their host and none of the cultures show evidence of multiple species of photobiont. This suggest that clonal photobiont samples exist in nature as samples such as the CCAP 1660/12 and 1660/13 cultures were sampled directly from the environment. As the Yad1g1N culture creation involved the isolation and purification of the 1N endosymbiont, clearing of the Yad1g host and then their subsequent reintroduction this suggests that the photobiont undergo little divergence and remain largely homogeneous within the host.

Unfortunately, the utility of the single cell metagenome to further test endosymbiont clonality was limited due to the high level of bacterial contamination. New tools and methods need to be developed to process and cluster MDA genomic contigs in the absence of reference genomes due to the unreliability of coverage as a feature. The existence of this type of method optimised for “de novo” assembled eukaryotic data would greatly aid the analysis of complex interacting eukaryotic systems using MDA based genomics. One approach could be to utilise blanket normalisation methods and base the variational inference off the normalised coverage and compositional features however, but using a blanket coverage threshold instead of a relative one there is the potential to discard a significant portion of sequencing data.

Therefore, the first result chapter already shed doubt on some of the arguments supporting the utility of *Paramecium* in the study of endosymbiosis. Specifically, it is not necessarily facultative and metabolic dependence if not necessarily co-dependence has become fixed in at least one species. Additionally, due to this evidence of diversity in the relationships combined with the taxonomic turmoil and previous mislabelling means that the utility of the reference literature is reduced. Any data from literature prior to the establishment of molecular taxonomy in these species needs to be carefully revisited and verified before it can be effectively used to contextualise the “omic” analyses.

In-depth analysis of the optimisation of pre-processing, filtering, assembly, and binning of single cell transcriptomics revealed that it was possible to generate and assemble single cell RNA-seq datasets of complex eukaryotic systems. Previous work has shown the potential utility of “de novo” SCT in eukaryotic micro-organisms (Kolisko et al., 2014) however, this work represents the first analysis using SCT to investigate a non-axenic eukaryotic system. Particularly, this is also the first analysis of a pair of interacting eukaryotic partners using single cell methods in the absence of reference genomes. This analysis identified that GC% based pre-assembly read partitioning is ineffectual for these datasets, but taxonomic screening is highly necessary to minimise the levels of bacterial contamination. This data also emphasises the utility of phylogenetically informed transcript binning processes instead of relying exclusively on naive top BLAST hit approaches. This work also determined that current recommended practices in bulk RNA-seq, such as digital normalisation and error correction, are still highly useful techniques in the analysis and assembly of SCT datasets. Future work could consider the utility of phylogenetically aware kernels (e.g. (Vert, 2002)) in the classification of transcript bins. Pre-assembly read partitioning should be revisited and the benefit of incorporating additional sequence feature such as composition and coverage data investigated in this form a pre-processing.

An analysis of the endosymbiont metabolism via expressed transporters and secreted proteins revealed novel aspects of amino acid usage by *M. reissieri* as well as the potential synthesis of complex saccharides such as raffinose and arabinose within the PV lumen. Metabolomic data supporting these hypotheses were also presented. While the untargeted metabolomic profiling does require further optimisation, particularly GC-QTOF analysis of carbohydrate metabolism, these approaches are highly useful to supplement transcriptomic data. Further, targeted mass spectrometry is required to confirm the differential abundances of raffinose, arabinose. Additionally, the targeted amino acid analysis requires redone due to a failure to fit calibration curves to the peaks generated by the majority of the amino acids.

Comparison of the active metabolic network in *M. reissieri* during endosymbiosis to that of *C. variabilis* 1N and the total metabolic capacity of *C. variabilis* NC64A and *Coccomyxa subellipsoidea* also revealed unique traits. Specifically, *M. reissieri* does not express aspects of fatty acid degradation present in the other endosymbionts as well as having distinct amino acid degradation pathways that are congruent with the identified alternative amino acid usage in this species.

The discovery of novel traits support the potential utility for single cell transcriptomics and bulk metabolomic analysis for identifying the underlying molecular function of a given endosymbiotic relationship. Unfortunately, it also further underlines the diversity and variability displayed between different *P. bursaria* - green algal systems.

Finally, an analysis of RNAi in *P. bursaria* revealed a potentially inactive/absent dsRNA induced RNAi system in *P. bursaria* SW1 (CCAP 1660/12). The common pattern of presence and absence of the previously identified components of the RNAi pathways in both *P. bursaria* transcriptomes suggests that this system is likely to be inactive or missing in *P. bursaria*. The most significantly missing factor is that of the *Pds1* gene that has been implicated in the uptake of RNA from the digestive vacuole. As this has been discovered to occur at low and natural levels in *P. tetaurelia* during normal feeding (Carradec et al., 2015) the potentially deleterious presence of eukaryotic algal endosymbionts may offer an explanation for the deactivation/absence of this system.

An “in-silico” study of the number of potential siRNA “collisions” between the active *Paramecium* transcriptomes and other eukaryotic transcriptomes (particularly those belonging to the endosymbionts) supported this hypothesis. Relatively more collisions occur between the host and eukaryotic transcriptomes than do versus bacterial ones. The greater number of collisions increasing the chance of deleterious cross-talk taking place and thus likely increasing the fitness cost of maintaining this system in the presence of eukaryotic endosymbiont. Additional work needs done to assess the exact nature of these collisions, particularly between host and endosymbiont.

As there is evidence of both specialised adaptations in each host-algal system, including potentially mutually obligate dependencies, why is there a lack of evidence evidence of tighter integration in this system? Specifically, the type of genomic integration displayed in other endosymbionts such as EGT.

Firstly, the protein import systems considered necessary for extensive EGT to start taking place are complicated in the cases of secondary and tertiary endosymbioses than basic plastids due to the increased number of membranes that may need to be traversed especially for import directly to the secondary plastid from the host (Hirakawa et al., 2012). Secondly, the unusual nuclear dimorphism of the host *P. bursaria* and codon usage may prove a barrier to the vast majority of EGT activity.

For successful transfer to take place between host and endosymbiont it would be necessary for the gene to transfer not just from the endosymbiont to the transcriptionally active host MAC but to the germline MIC. Even then integration into the MIC would have to occur in such a way that it would be correctly spliced and duplicated during the conversion of the MIC back to the MAC. Compounding this with sexual reproduction further decreases the probability of effective integration. It is notable that the prototypical hosts of the endosymbiotically “promiscuous” green algae - *Chlorella*, *Coccomyxa* and *Micractinium* all display germline sequestration either through the aforementioned dimorphism in *P. bursaria* or via standard metazoan germlines in the case of *Hydra* (Kawaida et al., 2013) and the kleptoplastic sacoglossan sea slugs (Yellowlees et al., 2008).

Future work, could attempt to use the genomic contigs generated here and/or further sequencing to attempt to pin-point examples of endosymbiont genes being present in host contigs and vice versa. Then, due to the rate of chimeric contigs in MDA, PCR and sanger sequencing could be used to confirm any putative EGTS.

Another interesting angle of investigation of these systems is analysing what host and endosymbiont are not expressing during endosymbiosis. This was partially investigated by (Kodama et al., 2014) however, the partial genome and transcriptome could be used to further investigate this question. Specifically, all genes present in the assembly could be identified and annotated using standard annotation pipelines and then the transcriptomes surveyed for their presence. Any gene present in the genome that is not observed in the endosymbiosis transcri-

tomes could shed further light on the function and evolution of these systems.

Ultimately, this work has identified the diverse, complex and distinct set of traits that different *P. bursaria*-green algal endosymbioses and demonstrated that, while still nascent, single cell methodologies can be amenable to the analysis of complex multimember eukaryotic systems that lack prior genomic references. Unfortunately, both due to the diversity of the systems discovered in the *P. bursaria*-algal endosymbioses and the inability to induce RNAi the initial utility of the *Paramecium bursaria* as a general model for the evolution of co-dependence is less than initially believed. However, future work in these systems using the “omic” resources generated in this thesis as a base dataset could help us understand how such mechanistic endosymbiotic diversity is possible even in closely related host and endosymbionts species. Understanding the answer to this question would greatly improve our understanding of the evolution of endosymbiosis and ultimately the eukaryotic cell.

Bibliography

- Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105.
- Abou-Shanab, R. a. I., El-Dalatony, M. M., EL-Sheekh, M. M., Ji, M.-K., Salama, E.-S., Kabra, A. N., and Jeon, B.-H. (2014). Cultivation of a new microalga, *Micractinium reisseri*, in municipal wastewater for nutrient removal, biomass, lipid, and fatty acid production. *Biotechnol. Bioprocess Eng.*, 19:510–518.
- Abrams, P. (1987). On classifying interactions between populations. *Oecologia*, 73:272–281.
- Achilles-Day, U. E. and Day, J. G. (2013a). Isolation of clonal cultures of endosymbiotic green algae from their ciliate hosts. *J. Microbiol. Methods*, 92(3):355–357.
- Achilles-Day, U. E. M. and Day, J. G. (2013b). Isolation of clonal cultures of endosymbiotic green algae from their ciliate hosts. *J. Microbiol. Methods*, 92(3):355–7.
- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, a., Olde, B., and Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(1990):1651–1656.
- Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J. J., Mayer, P., and Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.*, 28(20):E87.
- Adl, S., Simpson, A., Lane, C., Lukeš, J., Bass, D., Bowser, S., Brown, M., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., LeGall, L., Lynn, D., McManus, H., Mitchell, E., Mozley-Stanridge, S., Parfrey, L., Pawłowski, J., Rueckert, S., Shadwick, L., Schoch, C., Smirnov, A., and Spiegel, F. (2013). The Revised Classification of Eukaryotes. *J. Eukaryot. Microbiol.*, 59(5):1–45.
- Agre, P., Preston, G. M., Smith, B. L., Jung, J. S., Raina, S., Moon, C., Guggino, W. B., and Nielsen, S. (1993). Aquaporin CHIP: the archetypal molecular water channel. *Am. J. Physiol.*, 265:F463–F476.
- Aho, A. V., Kernighan, B. W., and Weinberger, P. J. (1987). *The AWK Programming Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19:716–723.
- Albers, D., Reisser, W., and Wiessner, W. (1982). Studies on the nitrogen supply of endosymbiotic chlorellae in green paramecium bursaria. *Plant Sci. Lett.*, 25:85–90.
- Alberts, B. (2015). *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, New York, NY.
- Aliaga Goltsman, D. S., Comolli, L. R., Thomas, B. C., and Banfield, J. F. (2014). Community transcriptomics reveals unexpected high microbial diversity in acidophilic biofilm communities. *ISME J.*, 9(4):1014–1023.
- Alibu, V. P., Storm, L., Haile, S., Clayton, C., and Horn, D. (2005). A doubly inducible system for RNA interference and rapid RNAi plasmid construction in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.*, 139:75–82.
- Allaire, J. J., Horner, J., Marti, V., and Porte, N. (2014). *markdown: Markdown rendering for R*.
- Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 11(11):1144–1146.
- Álvarez Loayza, P., White, J. F., Torres, M. S., Balslev, H., Kristiansen, T., Svenning, J. C., and Gil, N. (2011). Light converts endosymbiotic fungus to pathogen, influencing seedling survival and niche-space filling of a common tropical tree, *Iriartea deltoidea*. *PLoS One*, 6(1).
- Andrews, S. (2015). FastQC: A Quality Control tool for High Throughput Sequence Data.
- Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.*, 55(4):539–52.
- Antonov, A. V., Dietmann, S., and Mewes, H. W. (2008). KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, 9(12):R179.
- Aoki, K., Moriguchi, H., Yoshioka, T., Okawa, K., and Tabara, H. (2007). In vitro analyses of the production and activity of secondary small interfering RNAs in *C. elegans*. *EMBO J.*, 26(24):5007–5019.
- Archibald, J. (2005). Jumping Genes and Shrinking Genomes – Probing the Evolution of Eukaryotic Photosynthesis with Genomics. *IUBMB Life (International Union Biochem. Mol. Biol. Life)*, 57(August):539–547.
- Archibald, J. (2015). Endosymbiosis and Eukaryotic Cell Evolution. *Curr. Biol.*, 25(19):R911–R921.
- Archibald, J. M. (2014). *One plus one equals one: Symbiosis and the evolution of complex life*. Oxford University Press.
- Arnaiz, O., Cain, S., Cohen, J., and Sperling, L. (2007). ParameciumDB: A community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Res.*, 35(November 2006):439–444.

- Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E., and Sperling, L. (2010). Gene expression in a paleopolyploid: a transcriptome resource for the ciliate Paramecium tetraurelia. *BMC Genomics*, 11:547.
- Arnaiz, O. and Sperling, L. (2011a). ParameciumDB in 2011: New tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. *Nucleic Acids Res.*, 39(October 2010):632–636.
- Arnaiz, O. and Sperling, L. (2011b). ParameciumDB in 2011: New tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. *Nucleic Acids Res.*, 39(October 2010):632–636.
- Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *Open Bioinforma. J.*, 7:1–8.
- Arp, a. J., Childress, J. J., Arp, a. J., Childress, J. J., Childress, J. J., Fisher, C. R., Girguis, P. R., Childress, J. J., Akashi, H., Gojobori, T., Rau, G. H., Hedges, J. I., Rau, G. H., Fisher, C. R., and Brooks, J. M. (2007). Amplified Silencing. 315(January):199–200.
- Ashburner, M., Ball, C. a., Blake, J. a., Botstein, D., Butler, H., Cherry, J. M., Davis, a. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. a., Hill, D. P., Issel-Tarver, L., Kasarskis, a., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(may):25–29.
- Attwood, T. K., Beck, M. E., Bleasby, a. J., and Parry-Smith, D. J. (1994). PRINTS—a database of protein motif fingerprints. *Nucleic Acids Res.*, 22(17):3590–3596.
- Aubusson-Fleury, A., Cohen, J., and Lemullois, M. (2015). Chapter 22 - Ciliary heterogeneity within a single cell: The Paramecium model. In Basto, R. and Marshall, W. F., editors, *Methods in Cilia & Flagella*, volume 127 of *Methods in Cell Biology*, pages 457–485. Academic Press.
- Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–16.
- Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouél, A., Lepère, G., Mialis, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulaïn, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., and Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, 444(7116):171–8.
- Bachvaroff, T. R., Handy, S. M., Place, A. R., and Delwiche, C. F. (2011). Alveolate phylogeny inferred using concatenated ribosomal proteins. *J. Eukaryot. Microbiol.*, 58(3):223–33.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., and Pevzner,

- P. a. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.*, 19(5):455–477.
- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305.
- Barker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S. L., Ledley, R. S., Mewes, H. W., Pfeiffer, F., and Tsugita, a. (1998). The PIR-international Protein Sequence Database. *Nucleic Acids Res.*, 26(1):27–32.
- Bashan, Y., Lopez, B. R., Huss, V. a. R., Amavizca, E., and De-Bashan, L. E. (2015). Chlorella sorokiniana (formerly C. vulgaris) UTEX 2714, a non-thermotolerant microalga useful for biotechnological applications and as a reference strain. *J. Appl. Phycol.*
- Bateman, a. (2002). The Pfam Protein Families Database. *Nucleic Acids Res.*, 30(1):276–280.
- Batista, T. M. and Marques, J. a. T. (2011). RNAi pathways in parasitic protists and worms. *J. Proteomics*, 74(9):1504–1514.
- Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., and Bétermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. *Genes Dev.*, 23:2478–2483.
- Beisson, J., Bétermier, M., Bré, M. H., Cohen, J., Duharcourt, S., Duret, L., Kung, C., Malinsky, S., Meyer, E., Preer, J. R., and Sperling, L. (2010a). DNA microinjection into the macronucleus of paramecium. *Cold Spring Harb. Protoc.*, 5(1):1–4.
- Beisson, J., Bétermier, M., Bré, M.-H., Cohen, J., Duharcourt, S., Duret, L., Kung, C., Malinsky, S., Meyer, E., Preer, J. R., and Sperling, L. (2010b). Silencing specific Paramecium tetraurelia genes by feeding double-stranded RNA. *CSH Protoc.*, 2010(1):pdb.prot5363.
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Anal. ...*, (1993):1–30.
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy Layer-Wise Training of Deep Networks. *Adv. Neural Inf. Process. Syst.*, 19(1):153.
- Bengtsson, M., Ståhlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. pages 1388–1392.
- Berger, J. (1980). Feeding Behaviour of Didinium nasutum on Paramecium bursaria with Normal or Apochlorotic Zoochlorellae. *J. Gen. Microbiol.*, 118:397–404.
- Berger, J. D. (1986). Autogamy Cell Cycle Stage-specific in Paramecium Commitment to Meiosis. *Exp. Cell Res.*, 166:475–485.

- Bergsten, J. (2005). A review of long-branch attraction. *21*:163–193.
- Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *J. of Machine Learn. Res.*, *13*:281–305.
- Berk, S. G., Parks, L. H., and Ting, R. S. (1991). Photoadaptation alters the ingestion rate of *Paramecium bursaria*, a mixotrophic ciliate. *Appl. Environ. Microbiol.*, *57*:2312–2316.
- Bernstein, E., Caudy, a. a., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, *409*(1997):363–366.
- Bétermier, M. (2004). Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate *Paramecium*. *Res. Microbiol.*, *155*:399–408.
- Bhattacharya, D., Archibald, J. M., Weber, A. P. M., and Reyes-Prieto, A. (2007). How do endosymbionts become organelles? Understanding early events in plastid evolution. *Bioessays*, *29*:1239–46.
- Bianchi, L. and Díez-Sampedro, A. (2010). A Single Amino Acid Change Converts the Sugar Sensor SGLT₃ into a Sugar Transporter. *PLoS One*, *5*(4):e10241.
- Binga, E. K., Lasken, R. S., and Neufeld, J. D. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.*, *2*:233–241.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.
- Blainey, P. C. (2013). The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.*, *37*:407–427.
- Blainey, P. C. and Quake, S. R. (2011). Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.*, *39*(4):1–9.
- Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D. D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., Pangilinan, J., Pröschold, T., Salamov, A., Schmutz, J., Weeks, D., Yamada, T., Lomsadze, A., Borodovsky, M., Claverie, J.-M., Grigoriev, I. V., and Van Etten, J. L. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.*, *13*(5):R39.
- Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., Salamov, A., Terry, A., Yamada, T., Dunigan, D. D., Grigoriev, I. V., Claverie, J.-M., and Van Etten, J. L. (2010). The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell*, *22*(9):2943–55.
- Boenigk, J., Ereshefsky, M., Hoef-Emden, K., Mallet, J., and Bass, D. (2012). Concepts in protistology: Species definitions and boundaries. *Eur. J. Protistol.*, *48*(2):96–102.

- Boetius, a., Ravenschlag, K., Schubert, C. J., Rickert, D., Widdel, F., Gieseke, a., Amann, R., Jørgensen, B. B., Witte, U., and Pfannkuche, O. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature*, 407(October):623–626.
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics*, 30(15):2114–2120.
- Bonetta, L. (2006). Genome sequencing in the fast lane. *Nat. Methods*, 3(2):141–147.
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proc. 5th Annu. ACM Work. Comput. Learn. Theory*, pages 144–152.
- Bouchard-Côté, A. and Jordan, M. I. (2013). Evolutionary inference via the Poisson Indel Process. *Proc. Natl. Acad. Sci. U. S. A.*, 110(4):1160–6.
- Bouhouche, K., Gout, J.-F., Kapusta, a., Betermier, M., and Meyer, E. (2011). Functional specialization of Piwi proteins in Paramecium tetraurelia from post-transcriptional gene silencing to genome remodelling. *Nucleic Acids Res.*, 39(10):4249–4264.
- Bradley, P. S. and Bradley, P. S. (1998). Refining Initial Points for K-Means Clustering. *Microsoft Res.*, pages 91–99.
- Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. a., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. a., Ganapathy, G., Gibbs, R. a., Gnerre, S., Godzariadis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2:10.
- Bradski, G. (2000). OpenCV library. *Dr. Dobb's J. Softw. Tools*.
- Braslavsky, I., Braslavsky, I., Hebert, B., Hebert, B., Kartalov, E., Kartalov, E., Quake, S. R., and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.*, 100:3960–4.
- Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2015). Near-optimal RNA-Seq quantification. *arXiv*.
- Breiman, L. (2001). Random forests. *Mach. Learn.*, pages 5–32.
- Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv*, 1203.4802:1–18.

- Brown, J. a. and Nielsen, P. J. (1974). Transfer of photosynthetically produced carbohydrate from endosymbiotic chlorellae to Paramecium bursaria. *J. Protozool.*, 21:569–570.
- Bruno, W. J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, 13(10):1368–1374.
- Buchan, D. W. a., Shepherd, A. J., Lee, D., Pearl, F. M. G., Rison, S. C. G., Thornton, J. M., and Orengo, C. a. (2002). Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.*, 12:503–14.
- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and Sensitive Protein Alignment using DIAMOND. *Under Rev.*, 12(1).
- Buchheim, M. a., Keller, A., Koetschan, C., Förster, F., Merget, B., and Wolf, M. (2011). Internal transcribed spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: towards an automated reconstruction of the green algal tree of life. *PLoS One*, 6(2):e16931.
- Buchon, N. and Vaury, C. (2006). RNAi: a defensive RNA-silencing against viruses and transposable elements. *Heredity (Edinb)*, 96:195–202.
- Buck, A. H., Coakley, G., Simbari, F., McSorley, H. J., Quintana, J. F., Le Bihan, T., Kumar, S., Abreu-Goodger, C., Lear, M., Harcus, Y., Ceroni, A., Babayan, S. a., Blaxter, M., Ivens, A., and Maizels, R. M. (2014). Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nat. Commun.*, 5:5488.
- Burki, F. (2014). The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.*, 6(5):1–18.
- Camin, J. H. and Sokal, R. R. (1965). A Method for deducing branching sequences in phylogeny. *Evolution (N. Y.)*, 19(3):311–326.
- Camoni, L., Marra, M., Garufi, A., Visconti, S., and Aducci, P. (2006). The maize root plasma membrane H+-ATPase is regulated by a sugar-induced transduction pathway. *Plant Cell Physiol.*, 47(6):743–747.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.
- Carradec, Q., Gotz, U., Arnaiz, O., Pouch, J., Simon, M., Meyer, E., and Marker, S. (2015). Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate Paramecium tetraurelia. *Nucleic Acids Res.*, 43(3):1818–1833.
- Carthew, R. W. and Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655.
- Casadevall, A. and Fang, F. C. (2008). Descriptive Science. *Infect. Immun.*, 76(9):3835–3836.

- Caspi, R., Foerster, H., Fulcher, C. a., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, a. G., Tissier, C., Walk, T. C., Zhang, P., and Karp, P. D. (2007). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 36(October 2007):D623–D631.
- Cavalli-Sforza, L. L. and Edwards, a. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.*, 19(012):233–257.
- CCAP (2012). NCL Media Recipe.
- Cerutti, H. and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.*, 50:81–99.
- Chalker, D. L., Meyer, E., and Mochizuki, K. (2013). Epigenetics of ciliates. *Cold Spring Harb. Perspectives Epigenetics*, 5:a017764.
- Chambouvet, A., Gower, D. J., Jirk, M., Yabsley, M. J., Davis, A. K., Leonard, G., Maguire, F., Doherty-bone, T. M., Bittencourt-silva, G. B., and Wilkinson, M. (2015). Cryptic infection of a broad taxonomic and geographic diversity of tadpoles by Perkinsea protists. pages 1–9.
- Chang, S.-S., Zhang, Z., and Liu, Y. (2012). RNA Interference Pathways in Fungi: Mechanisms and Functions. *Annu. Rev. Microbiol.*, 66:305–323.
- Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C. L., and Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.*, 16:30.
- Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., and Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One*, 5(1):e8613.
- Chen, T.-T. (1940). Polyploidy in Paramecium bursaria. *Genetics*, 26:239–240.
- Cho, B. H., Sauer, N., Komor, E., and Tanner, W. (1981). Glucose induces two amino acid transport systems in Chlorella. *Proc. Natl. Acad. Sci. U. S. A.*, 78(6):3591–4.
- Cho, K.-O., Kim, G.-W., and Lee, O.-K. (2011). Wolbachia Bacteria Reside in Host Golgi-Related Vesicles Whose Position Is Regulated by Polarity Proteins. *PLoS One*, 6(7):e22703.
- Clavijo, B., Mapleson, D., Ayling, S., and Caccamo, M. (2015). KAT—Kmer Analysis Tool. Available online at: <http://www.tgac.ac.uk/kat>.
- Coleman, A. W. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.*, 19(7):370–375.
- Collins, K. and Lee, S. R. (2006). Two classes of endogenous small RNAs in Tetrahymena thermophila. *Genes Dev.*, 20:28–33.

- Conesa, a., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.
- Cooper, G. (2013). *The cell : a molecular approach*. Sinauer Associates, Sunderland, MA.
- Corduneanu, A. and Bishop, C. M. (2001). Variational Bayesian Model Selection for Mixture Distributions. *Artif. Intell.*, 51:27–34.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20:273–297.
- Crespi, M. and Frugier, F. (2008). De novo organ formation from differentiated cells: root nodule organogenesis. *Sci. Signal.*, 1(49):re11.
- Crick, F. (1966). *Of molecules and man*. University of Washington.
- Criscuolo, A. (2011). Molecular Phylogenetics and Evolution morePhyML : Improving the phylogenetic tree space exploration with PhyML 3. *Mol. Phylogenet. Evol.*, 61(3):944–948.
- Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., Gupta, A., Herr, J. R., Howe, A., Hyer, A., Härpfer, A., Irber, L., Kidd, R., Lin, D., Lippi, J., Mansour, T., McA'Nulty, P., McDonald, E., Mizzi, J., Murray, K. D., Nahum, J. R., Nanholy, K., Nederbragt, A. J., Ortiz-Zuazaga, H., Ory, J., Pell, J., Pepe-Ranney, C., Russ, Z. N., Schwarz, E., Scott, C., Seaman, J., Sievert, S., Simpson, J., Skennerton, C. T., Spencer, J., Srinivasan, R., Standage, D., Stapleton, J. a., Steinman, S. R., Stein, J., Taylor, B., Trimble, W., Wiencko, H. L., Wright, M., Wyss, B., Zhang, Q., Zyme, E., and Brown, C. T. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research*, pages 1–9.
- Cullis, C. A. (1972). DNA amounts in the nuclei of Paramecium bursaria. *Chromosoma*, 40:127–133.
- Curtin, R. R., Cline, J. R., Slagle, N. P., March, W. B., Ram, P., Mehta, N. A., and Gray, A. G. (2013). MLPACK: A Scalable C++ Machine Learning Library. *J. Mach. Learn. Res.*, 14:801–805.
- Cuvelier, M. L., Allen, a. E., Monier, a., McCrow, J. P., Messie, M., Tringe, S. G., Woyke, T., Welsh, R. M., Ishoey, T., Lee, J.-H., Binder, B. J., DuPont, C. L., Latasa, M., Guigand, C., Buck, K. R., Hilton, J., Thiagarajan, M., Caler, E., Read, B., Lasken, R. S., Chavez, F. P., and Worden, a. Z. (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci.*, 107(33):14679–14684.
- Damaj, R., Pomel, S., Bricheux, G., Coffe, G., Viguès, B., Ravet, V., and Bouchard, P. (2009). Cross-study analysis of genomic data defines the ciliate multigenic epiplasmin family: strategies for functional analysis in Paramecium tetraurelia. *BMC Evol. Biol.*, 9:125.
- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–5.

- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, 9(8):772–772.
- Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A Model of Evolutionary Change in Proteins.
- de Bary, A. (1869). *Dir Erscheinung der Symbiose*. Verlag von Karl J. Trübner, Strassburg.
- De Queiroz, K. (2007). Species concepts and species delimitation. *Syst. Biol.*, 56(6):879–886.
- Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS One*, 8(12):1–13.
- Delano, W. L. (2002). The PyMOL Molecular Graphics System.
- Dettmer, K., Aronov, P. a., and Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.*, 26(1):51–78.
- Ding, Y., Zhao, Y., Xipeng Shen, N., Musuvathi, M., and Todd Mytkowicz, M. (2015). Yinyang K-Means: A Drop-In Replacement of the Classic K-Means with Consistent Speedup. 37.
- Dohra, H., Fujishima, M., and Ishikawa, H. (1998). Structure and expression of a GroE-homologous operon of a macronucleus-specific symbiont Holospora obtusa of the ciliate Paramecium caudatum. *J. Eukaryot. Microbiol.*, 45:71–79.
- Dolan, J. (1992). Mixotrophy in Ciliates: A Review of Chlorella Symbiosis and Chloroplast Retention. *Mar. Microb. Food Webs*, 6(2):115–132.
- Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11.
- Dorrell, R. G. and Smith, A. G. (2011). Do red and green make brown?: Perspectives on plastid acquisitions within chromalveolates. *Eukaryot. Cell*, 10(7):856–868.
- Dougherty, E. R. (2008). On the epistemological crisis in genomics. *Curr. Genomics*, 9:69–79.
- Douglas, a. E. H. V. a. R. (1986). On the characteristics and taxonomic position of symbiotic Chlorella. *Arch Microbiol*, 145:80–84.
- Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U. S. A.*, 100(M):8817–8822.
- Dröge, J. and McHardy, A. C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.*, 13(6):646–655.

- Duret, L., Cohen, J., Jubin, C., Dessen, P., Goût, J.-f., Mousset, S., Jaillon, O., Noël, B., Arnaiz, O., Bétermier, M., Wincker, P., Meyer, E., and Aury, J.-m. (2008). Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: A somatic view of the germline. *Genome Res.*, 18:585–596.
- Eck, R. V. and Dayhoff, M. O. (1966). Atlas of Protein Sequence and Structure.
- Eckert, K. and Kunkel, T. (1990). High fidelity DNA synthesis by the *Thermus aquaticus* polymerase. *Biochemistry*, 18(13):3739–3744.
- Eddy, S. and R, E. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, 7(10):e1002195.
- Eddy, S. R. (1995). Multiple alignment using hidden Markov model. *Intell. Syst. Mol. Biol.*, ISMB-95:114–120.
- Edgar, R. C. (2004a). MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(113):1–19.
- Edgar, R. C. (2004b). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.
- Eggensperger, K., Feurer, M., and Hutter, F. (2013). Towards an empirical foundation for assessing bayesian optimization of hyperparameters. ... *Bayesian Optim.* ..., pages 1–5.
- Eisen, J. a., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C. C., Gorovsky, M. a., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. a., Krieger, C. J., Del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. a., Hamilton, E. P., and Orias, E. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, 4(9):1620–1642.
- El Hadidi, M., Ruscheweyh, H.-J., and Huson, D. (2013). Improved metagenome analysis using MEGAN5. In *Jt. 21st Annu. Int. Conf. Intell. Syst. Mol. Biol. 12th Eur. Conf. Comput. Biol. 2013*.
- Elbashir, S. M., Martinez, J., Patkaniowska, a., Lendeckel, W., and Tuschl, T. (2001). Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.*, 20(23):6877–6888.
- Ellegaard, K. M., Klasson, L., and Andersson, S. G. E. (2013a). Testing the reproducibility of multiple displacement amplification on genomes of clonal endosymbiont populations. *PLoS One*, 8(11):21–25.
- Ellegaard, K. M., Klasson, L., Näslund, K., Bourtzis, K., and Andersson, S. G. E. (2013b). Comparative genomics of Wolbachia and the bacterial species concept. *PLoS Genet.*, 9(4):e1003381.
- Ender, C. and Meister, G. (2010). Argonaute proteins at a glance. *J. Cell Sci.*, 123:1819–1823.

- Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, pages 226–231.
- Fama, P., Olsen, J. L., Stam, W. T., and Procaccini, G. (2000). High levels of intra- and inter-individual polymorphism in the rDNA ITS1 of *Caulerpa racemosa* (Chlorophyta). *Eur. J. Phycol.*, 35(December):349–356.
- Fang, F. C. and Casadevall, A. (2011). Reductionistic and holistic science. *Infect. Immun.*, 79(4):1401–1404.
- Fayyad, U., Grinstein, G. G., and Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Res.*, 40(December 2011):136–143.
- Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34(3).
- Feldmesser, E., Rosenwasser, S., Vardi, A., and Ben-Dor, S. (2014). Improving transcriptome construction in non-model organisms: integrating manual and automated gene definition in *Emiliania huxleyi*. *BMC Genomics*, 15(1):148.
- Felsenstein, J. (1978). Society of Systematic Biologists. *Soc. Syst. Biol.*, 27(4):401–410.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.
- Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution (N. Y.)*, 39(4):783–791.
- Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Syst. Biol.*, 50(4):465–467.
- Fenchel, T. and Finlay, B. (1992). Production of methane and hydrogen by anaerobic ciliates containing symbiotic *Microbiology*, pages 475–480.
- Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25:351–360.
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.*, 15:3133–3181.
- Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48:155–171.
- Fire, a., Xu, S., Montgomery, M. K., Kostas, S. a., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(February):806–811.
- Fitch, W. M. and Magoliash, E. (1967). Construction of Phylogenetic Tress. *Science (80-.)*, 155(279).

- Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., Vlahovicek, K., and Svoboda, P. (2013). A retrotransposon-driven dicer isoform directs endogenous small interfering rna production in mouse oocytes. *Cell*, 155(4):807–816.
- Fokin, S. I. and Görtz, H.-d. (2009). Diversity of Holospora Bacteria in Paramecium and Their Characterization. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 7, pages 161–199. Springer.
- Fokin, S. I., Przyboś, E., Chivilev, S. M., Beier, C. L., Horn, M., Skotarczak, B., Wodecka, B., and Fujishima, M. (2004). Morphological and molecular investigations of Paramecium schewiakoffi sp. nov. (Ciliophora, Oligohymenophorea) and current status of distribution and taxonomy of Paramecium spp. *Eur. J. Protistol.*, 40:225–243.
- Forgy, E. W. (1965). Cluster Analysis of Multivariate Data : Efficiency Versus Interpretability of Classifications. *Biometrics*, 21:768–769.
- Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55:119–139.
- Funfak, A., Fisch, C., Abdel Motaal, H. T., Diener, J., Combettes, L., Baroud, C. N., and Dupuis-Williams, P. (2015). Paramecium swimming and ciliary beating patterns: a study on four RNA interference mutations. *Integr. Biol.*, 7:90–100.
- Gabow, H. N. and Tarjan, R. E. (1985). A linear-time algorithm for a special case of disjoint set union. *J. Comput. Syst. Sci.*, 30:209–221.
- Gabrielsen, T. M., Minge, M. a., Espelund, M., Tooming-Klunderud, A., Patil, V., Nederbragt, A. J., Otis, C., Turmel, M., Shalchian-Tabrizi, K., Lemieux, C., and Jakobsen, K. S. (2011). Genome evolution of a tertiary dinoflagellate plastid. *PLoS One*, 6(4).
- Galluzzi, L., Penna, A., Bertozzini, E., Garcés, E., Magnani, M., Vila, M., and Garce, E. (2004). Development of a Real-Time PCR Assay for Rapid Detection and Quantification of *Alexandrium minutum* (a Dinoflagellate) Development of a Real-Time PCR Assay for Rapid Detection and Quantification of *Alexandrium minutum* (a Dinoflagellate). *Appl. Environ. Microbiol.*, 70(2):1199–1206.
- Galvani, A. and Sperling, L. (2002). RNA interference by feeding in Paramecium. *Trends Genet.*, 18(1):11–2.
- Galvani, A., Sperling, L., Moléculaire, C. D. G., and Cedex, G.-s.-y. (2001). Transgene-mediated post-transcriptional gene silencing is inhibited by 3' non-coding sequences in Paramecium. *Nucleic Acids Res.*, 29(21):4387–4394.
- Garcia-Cuetos, L., Moestrup, O., and Hansen, P. J. (2012). Studies on the genus mesodinium II. Ultrastructural and molecular investigations of five marine species help clarifying the taxonomy. *J. Eukaryot. Microbiol.*, 59(4):374–400.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14:685–695.

Gelman, a., Hill, J., and Yajima, M. (2009). Why we (usually) don't have to worry about multiple comparisons. *arXiv:0907.2478 [stat.AP]*, pages 1–25.

Gene Codes (2015). Sequencher® version 4.10.1 sequence analysis software.

Gerhard, D. S., Wagner, L., Feingold, E. a., Shenmen, C. M., Grouse, L. H., Schuler, G., Klein, S. L., Old, S., Rasooly, R., Good, P., Guyer, M., Peck, A. M., Derge, J. G., Lipman, D., Collins, F. S., Jang, W., Sherry, S., Feolo, M., Misquitta, L., Lee, E., Rotmistrovsky, K., Greenhut, S. F., Schaefer, C. F., Buetow, K. H., Bonner, T. I., Haussler, D., Kent, J., Diekhans, M., Furey, T., Brent, M., Prange, C., Schreiber, K., Shapiro, N., Bhat, N. K., Hopkins, R. F., Hsie, F., Driscoll, T., Soares, M. B., Bonaldo, M. F., Casavant, T. L., Scheetz, T. E., Brownstein, M. J., Usdin, T. B., Toshiyuki, S., Carninci, P., Piao, Y., Dudekula, D. B., Ko, M. S. H., Kawakami, K., Suzuki, Y., Sugano, S., Gruber, C. E., Smith, M. R., Simmons, B., Moore, T., Waterman, R., Johnson, S. L., Ruan, Y., Lin Wei, C., Mathavan, S., Gunaratne, P. H., Wu, J., Garcia, A. M., Hulyk, S. W., Fuh, E., Yuan, Y., Sneed, A., Kowis, C., Hodgson, A., Muzny, D. M., McPherson, J., Gibbs, R. a., Fahey, J., Helton, E., Ketteman, M., Madan, A., Rodrigues, S., Sanchez, A., Whiting, M., Madan, A., Young, A. C., Wetherby, K. D., Granite, S. J., Kwong, P. N., Brinkley, C. P., Pearson, R. L., Bouffard, G. G., Blakesly, R. W., Green, E. D., Dickson, M. C., Rodriguez, A. C., Grimwood, J., Schmutz, J., Myers, R. M., Butterfield, Y. S. N., Griffith, M., Griffith, O. L., Krzywinski, M. I., Liao, N., Morrin, R., Palmquist, D., Petrescu, A. S., Skalska, U., Smailus, D. E., Stott, J. M., Schnurch, A., Schein, J. E., Jones, S. J. M., Holt, R. a., Baross, A., Marra, M. a., Clifton, S., Makowski, K. a., Bosak, S., and Malek, J. (2004). The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.*, 14:2121–2127.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.*, 63(October 2005):3–42.

Gilbert, D. G. (2013). Gene-omes built from mRNA-seq not genome DNA. In *7th Annu. Arthropod Genomics Symp*, Notre Dame.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.*, 11:759–769.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, 40(November 2011):D1178–D1186.

Gordon, A. and Hannon, G. J. (2010). Fastx-toolkit. *FASTQ/A Short-reads Pre-processing Tools (Unpublished)* http://hannonlab.cshl.edu/fastx_toolkit.

Gortz, H.-d. (1982). Infections of Paramecium Bursaria with Bacteria and Yeasts. *J. Cell Sci.*, 58:445–453.

- Görtz, H.-d. and Fokin, S. I. (2009). Diversity of Endosymbiotic Bacteria in Paramecium. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 6, pages 131–160. Springer.
- Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, 30(1):268–72.
- Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, 27(2):221–4.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7).
- Green, B. R. (2011). After the primary endosymbiosis: An update on the chromalveolate hypothesis and the origins of algae with Chl c. *Photosynth. Res.*, 107:103–115.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2015). LSTM: A Search Space Odyssey. *arXiv*, page 10.
- Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., and Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat. Rev. Microbiol.*, 8(10):743–52.
- Gribskov, M., Homyak, M., Edenfield, J., and Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences.
- Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R. a., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D., and Dubchak, I. (2011). P@JGI-DOE@The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.*, 40(November 2011):1–7.
- Guarnieri, M. T., Nag, A., Smolinski, S. L., Darzins, A., Seibert, M., and Pienkos, P. T. (2011). Examination of triacylglycerol biosynthetic pathways via de novo transcriptomic and proteomic analyses in an unsequenced microalga. *PLoS One*, 6(10).
- Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.*, 52(5):696–704.
- Gunderson, J. H., Elwood, H., Ingold, a., Kindle, K., and Sogin, M. L. (1987). Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. *Proc Natl Acad Sci U S A*, 84(August):5823–5827.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., Leduc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8(8):1494–512.
- Haft, D. H. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.*, 31(1):371–373.
- Hamel, A., Fisch, C., Combettes, L., Dupuis-Williams, P., and Baroud, C. N. (2011). Transitions between three swimming gaits in Paramecium escape. *Proc. Natl. Acad. Sci. U. S. A.*, 108:7290–7295.
- Harris, M. a., Clark, J., Ireland, a., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. a., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, a., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32:D258–D261.
- Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, 320(April):106–109.
- Hasegawa, M., Kishino, H., and Yano, T. A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174.
- Hayden, K. J., Garbelotto, M., Knaus, B. J., Cronn, R. C., Rai, H., and Wright, J. W. (2014). Dual RNA-seq of the plant pathogen Phytophthora ramorum and its tanoak host. *Tree Genet. Genomes*, 10:489–502.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:1502.01852*.
- Heeg, J. S. and Wolf, M. (2015). ITS₂ and 18S rDNA sequence-structure phylogeny of Chlorella and allies (Chlorophyta, Trebouxiophyceae, Chlorellaceae). *Plant Gene*, 4:20–28.
- Helber, N., Wippel, K., Sauer, N., Schaarschmidt, S., Hause, B., and Requena, N. (2011). A Versatile Monosaccharide Transporter That Operates in the Arbuscular Mycorrhizal Fungus Glomus sp Is Crucial for the Symbiotic Relationship with Plants. *Plant Cell*, 23(October):3812–3823.
- Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89(November):10915–10919.

- Henneberger, R., Moissl, C., Amann, T., Rudolph, C., and Huber, R. (2006). New insights into the lifestyle of the cold-loving SM1 euryarchaeon: Natural growth as a monospecies biofilm in the subsurface. *Appl. Environ. Microbiol.*, 72(1):192–199.
- Hershkovitz, M. and Lewis, L. A. (1996). Deep-Level Diagnostic Value of the rDNA-ITS Region. *Mol Biol Evol*, 13(9):1276–1295.
- Hinton, G. and Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science* (80-.), 313(July):504–507.
- Hirakawa, Y., Burki, F., and Keeling, P. J. (2012). Genome-based reconstruction of the protein import machinery in the secondary plastid of a chlorarachniophyte alga. *Eukaryot. Cell*, 11:324–333.
- Hirsch, A. A. (1992). Developmental biology of legume nodulation. *New Phytol.*, (40):211–237.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1–32.
- Hodges, M. E., Wickstead, B., Gull, K., and Langdale, J. a. (2011). Conservation of ciliary proteins in plants with no cilia. *BMC Plant Biol.*, 11(1):185.
- Hoffman, P., Grinstein, G., Marx, K., Grosse, I., and Stanley, E. (1997). DNA visual and analytic data mining.
- Hoffmeister, M. and Martin, W. (2003). Interspecific evolution: microbial symbiosis, endosymbiosis and gene transfer. *Environ. Microbiol.*, 5:641–9.
- Hori, H., Lim, B. L., and Osawa, S. (1985). Evolution of green plants as deduced from 5S rRNA sequences. *Proc. Natl. Acad. Sci. U. S. A.*, 82(February):820–823.
- Hörtnagl, P. H. and Sommaruga, R. (2007). Photo-oxidative stress in symbiotic and aposymbiotic strains of the ciliate Paramecium bursaria. *Photochem. Photobiol. Sci.*, 6(8):842–847.
- Horton, P. and Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 5:147–152.
- Horton, P., Park, K., Obayashi, T., and Nakai, K. (2006). Protein Subcellular Localisation Prediction with WoLF PSORT. *Ser Adv Bioinform*, 3:39–48.
- Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, 35(17):W585–W587.
- Hoshina, R. (2012). Photobiont Flexibility in Paramecium bursaria: Double and Triple Photobiont Co-Habitation. *Adv. Microbiol.*, 02(September):227–233.
- Hoshina, R. and Fujiwara, Y. (2013). Molecular characterization of Chlorella cultures of the National Institute for Environmental Studies culture collection with description of Micractinium inermum sp. nov., Didymogenes sphaerica sp. nov., and Didymogenes soliella sp. nov. (Chlorellaceae, Tr. *Phycol. Res.*, 61(2):124–132.

- Hoshina, R. and Imamura, N. (2008). Multiple Origins of the Symbioses in *Paramecium bursaria*. *Protist*, 159 (January).
- Hoshina, R. and Imamura, N. (2009). Origins of Algal Symbionts of *Paramecium bursaria*. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs* 12, chapter 1, pages 2–29. Springer.
- Hoshina, R., Iwataki, M., and Imamura, N. (2010). Chlorella variabilis and Micractinium reisseri sp. nov. (Chlorophyceae, Trebouxiophyceae): Redescription of the endosymbiotic green algae of *Paramecium bursaria* (Peniculida, Oligohymenophorea) in the 120th year. *Phycol. Res.*, 58(3):188–201.
- Hoshina, R., Kamako, S.-I., and Imamura, N. (2004). Phylogenetic position of endosymbiotic green algae in *Paramecium bursaria* Ehrenberg from Japan. *Plant Biol. (Stuttg.)*, 6(4):447–53.
- Hosono, S., Faruqi, a. F., Dean, F. B., Du, Y., Sun, Z., Wu, X., Du, J., Kingsmore, S. F., Egholm, M., and Lasken, R. S. (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Res.*, 13:954–964.
- Hosoya, H., Kimura, K., Matsuda, S., Kitaura, M., Takahashi, T., and Kosaka, T. (1995). Symbiotic Algae-free Strains of The Green *Paramecium* *Paramecium bursaria* Produced by the Herbicide Paraquat. *Zoolog. Sci.*, 12:807–810.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417.
- Huber, H., Hohn, M. J., Rachel, R., and Fuchs, T. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, 417(May):63–67.
- Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–5.
- Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, 11:24.
- Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.*, 17:377–386.
- Huygens, C. (1899). *Œuvres complètes de Christiaan Huygens, Correspondances Volume VII: 1676-1684*. Dutch Society of Sciences, The Hague.
- ISO International Standard (2011). ISO/IEC 14882:2011(E) – Programming Language C++.
- Iwasaki, Y. W., Siomi, M. C., and Siomi, H. (2015). PIWI-Interacting RNA: Its Biogenesis and Functions. *Annu. Rev. Biochem.*, 84:405–433.
- Iwatsuki, K. and Naitoh, Y. (1988). Behavioural responses to light in *Paramecium bursaria* in relation to its symbiotic green alga Chlorella. *J. Exp. Biol.*, 60:43–60.

Jahn, C. L. and Klobutcher, L. a. (2002). Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.*, 56:489–

520.

Jaszczyszyn, Y., Thermes, C., and Dijk, E. L. V. (2014). Ten years of next-generation sequencing technology. *Trends Genet.*, 30(9).

Jennings, H. S. (1939). Genetics of Paramecium Bursaria. I. Mating Types and Groups, Their Interrelations and Distribution; Mating Behavior and Self Sterility. *Genetics*, 24(March):202–233.

Jiggins, C. D., Hirschmann, R. J., Bundey, R. a., Insel, P. a., Crossland, J. P., Pool, J. E., Kassner, V. a., Aquadro, C. F., Carroll, S. B., Swinehart, J., Kingsley, E. P., Jensen, J. D., Hoekstra, H. E., Larson, J. G., Manceau, M., Wiley, C. D., Stephens, M., Duhl, D. M. J., Millar, S. E., Miller, K. a., Barsh, G. S., Hernandez, R. D., Williamson, S. H., Bustamante, C. D., Kim, Y., Hartl, D. L., Parsch, J., Goncalves, G., Chupasko, J., Kay, E., Kingsley, E., Demboski, J., Foundation, S., Agency, P., Fellowship, P., and Commission, P. (2013). Circadian Control of Chloroplast. *Nature*, 339(March):1316–1319.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11:431.

Johnson, M. D. (2011). Acquired phototrophy in ciliates: a review of cellular interactions and structural adaptations. *J. Eukaryot. Microbiol.*, 58(3):185–95.

Jones, M., Dry, I. R., Frampton, D., Singh, M., Kanda, R. K., Yee, M. B., Kellam, P., Hollinshead, M., Kinchington, P. R., O'Toole, E. a., and Breuer, J. (2014). RNA-seq Analysis of Host and Viral Gene Expression Highlights Interaction between Varicella Zoster Virus and Keratinocyte Differentiation. *PLoS Pathog.*, 10(1).

Joshi, N. and Fass, J. (2015). Sickle: A Sliding-Window, Adaptive, Quality-based Trimming Tool for FastQ Files (Version 1.33).

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mamm. protein Metab.*, 3:21–132.

Kadono, T., Kawano, T., Hosoya, H., and Kosaka, T. (2004). Flow cytometric studies of the host-regulated cell cycle in algae symbiotic with green paramecium. *Protoplasma*, 223:133–141.

Kafsack, B. F. and Llinás, M. (2010). Eating at the Table of Another: Metabolomics of Host-Parasite Interactions. *Cell Host Microbe*, 7:90–99.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, 24:1384–1395.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res.*, 35:429–432.

- Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, 7(8):1511–1522.
- Kamako, S.-i., Hoshina, R., Ueno, S., and Imamura, N. (2005). Establishment of axenic endosymbiotic strains of Japanese Paramecium bursaria and the utilization of carbohydrate and nitrogen compounds by the isolated algae. *Eur. J. Protistol.*, 41:193–202.
- Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K., and Nakamura, Y. (2011). DDBJ progress report. *Nucleic Acids Res.*, 39(September 2010):D22–7.
- Kandler, O. and Hopf, H. (1982). Oligosaccharides Based on Sucrose (Sucrosyl Oligosaccharides). In Loewus, F. and Tanner, W., editors, *Plant Carbohydrates I SE - 8*, volume 13 / A of *Encyclopedia of Plant Physiology*, pages 348–383. Springer Berlin Heidelberg.
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(November 2013):D199–D205.
- Kapaun, E. and Reisser, W. (1995). A chitin-like glycan in the cell wall of a Chlorella sp. (Chlorococcales, Chlorophyceae). *Planta*, 197:577–582.
- Karakashian, S. (1963). Growth of Paramecium bursaria as Influenced by the Presence of Algal Symbionts. *Physiol. Zool.*, 36(1):52–68.
- Karkar, S., Facchinelli, F., Price, D. C., Weber, A. P. M., and Bhattacharya, D. (2015). Metabolic connectivity as a driver of host and endosymbiont integration. *Proc. Natl. Acad. Sci. U. S. A.*, page 201421375.
- Karp, P. D., Paley, S., and Romero, P. (2002). The Pathway Tools software. *Bioinformatics*, 18 Suppl 1:S225–S232.
- Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, a., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. (2010). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, 11(1):40–79.
- Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complex. Comput. Comput. Proc. a Symp. Complex. Comput. Comput.*, pages 85–103.
- Kato, Y. and Imamura, N. (2008). Effect of calcium ion on uptake of amino acids by symbiotic Chlorella F36-ZK isolated from Japanese Paramecium bursaria. *Plant Sci.*, 174(1):88–96.
- Kato, Y. and Imamura, N. (2009a). Amino acid transport systems of Japanese Paramecium symbiont F36-ZK. *Symbiosis*, pages 99–107.
- Kato, Y. and Imamura, N. (2009b). Metabolic Control Between the Symbiotic Chlorella and the Host Paramecium. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 3, pages 57–82. Springer.

- Kato, Y., Ueno, S., and Imamura, N. (2006). Studies on the nitrogen utilization of endosymbiotic algae isolated from Japanese Paramecium bursaria. *Plant Sci.*, 170(3):481–486.
- Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, 33(2):511–518.
- Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30(14):3059–3066.
- Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, 30(4):772–780.
- Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. In Dodge, Y., editor, *Stat. Data Anal. Based L 1-Norm Relat. Methods*, pages 405–416.
- Kawahara, Y., Oono, Y., Kanamori, H., Matsumoto, T., Itoh, T., and Minami, E. (2012). Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. *PLoS One*, 7(11):e49423.
- Kawaiida, H., Ohba, K., Koutake, Y., Shimizu, H., Tachida, H., and Kobayakawa, Y. (2013). Symbiosis between hydra and chlorella: molecular phylogenetic analysis and experimental study provide insight into its origin and evolution. *Mol. Phylogenet. Evol.*, 66(3):906–14.
- Keeling, P. J. (2004). Diversity and Evolutionary History of Plastids and Their Hosts the Tree of Eukaryotes. *Am. J. Bot.*, 91(10):1481–1493.
- Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 365:729–748.
- Keeling, P. J. (2013). The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu. Rev. Plant Biol.*, 64:583–607.
- Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*, 9(august):605–618.
- Keller, A., Schleicher, T., Förster, F., Ruderisch, B., Dandekar, T., Müller, T., and Wolf, M. (2008). ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales). *BMC Evol. Biol.*, 8:218.
- Keller, L. C., Romijn, E. P., Zamora, I., Yates, J. R., and Marshall, W. F. (2005). Proteomic Analysis of Isolated Chlamydomonas Centrioles Reveals Orthologs of Ciliary-Disease Genes. *Curr. Biol.*, 15:1090–1098.
- Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, 11(11):R116.
- Kessler, E. and Huss, V. A. R. (1990). Biochemical Taxonomy of Symbiotic Chlorella Strains from Paramecium and Acanthocystis*. *Bot. Acta*, 103(2):140–142.

- Ketting, R. F. (2011). The Many Faces of RNAi. *Dev. Cell*, 20:148–161.
- Kies, L. and Kremer, B. P. (1979). Function of cyanelles in the thecamoeba Paulinella chromatophora. *Naturwissenschaften*, 66:578–579.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36.
- Klukas, C. and Schreiber, F. (2007). Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, 23(3):344–350.
- Knip, M., Constantin, M. E., and Thordal-Christensen, H. (2014). Trans-kingdom cross-talk: small RNAs on the move. *PLoS Genet.*, 10(9):e1004602.
- Knittel, K. and Boetius, A. (2009). Anaerobic oxidation of methane: progress with an unknown process. *Annu. Rev. Microbiol.*, 63:311–334.
- Koch, a., Kumar, N., Weber, L., Keller, H., Imani, J., and Kogel, K.-H. (2013). Host-induced gene silencing of cytochrome P450 lanosterol C₁₄-demethylase-encoding genes confers strong resistance to Fusarium species. *Proc. Natl. Acad. Sci.*, 110(48):19324–19329.
- Kodama, Y. and Fujishima, M. (2008). Cycloheximide induces synchronous swelling of perialgal vacuoles enclosing symbiotic Chlorella vulgaris and digestion of the algae in the ciliate Paramecium bursaria. *Protist*, 159(May):483–94.
- Kodama, Y. and Fujishima, M. (2009). Infection of Paramecium bursaria by Symbiotic Chlorella Species. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 2, pages 31–55. Springer.
- Kodama, Y. and Fujishima, M. (2011). Endosymbiosis of Chlorella species to the ciliate Paramecium bursaria alters the distribution of the host's trichocysts beneath the host cell cortex. *Protoplasma*, 248:325–337.
- Kodama, Y. and Fujishima, M. (2012). Characteristics of the digestive vacuole membrane of the alga-bearing ciliate Paramecium bursaria. *Protist*, 163(4):658–70.
- Kodama, Y. and Fujishima, M. (2014). Symbiotic Chlorella variabilis incubated under constant dark conditions for 24 hours loses the ability to avoid digestion by host lysosomal enzymes in digestive vacuoles of host ciliate Paramecium bursaria. *FEMS Microbiol. Ecol.*, 90:946–955.
- Kodama, Y., Nakahara, M., and Fujishima, M. (2007). Symbiotic alga Chlorella vulgaris of the ciliate Paramecium bursaria shows temporary resistance to host lysosomal enzymes during the early infection process. *Protoplasma*, 230(1-2):61–7.
- Kodama, Y., Shumway, M., and Leinonen, R. (2012). The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, 40(October 2011):D54–D56.

- Kodama, Y., Suzuki, H., Dohra, H., Sugii, M., Kitazume, T., Yamaguchi, K., Shigenobu, S., and Fujishima, M. (2014). Comparison of gene expression of Paramecium bursaria with and without Chlorella variabilis symbionts. *BMC Genomics*, 15:183.
- Koga, R., Tsuchida, T., and Fukatsu, T. (2003). Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont Buchnera in an aphid. *Proc. Biol. Sci.*, 270(April):2543–2550.
- Kolisko, M., Boscaro, V., Burki, F., Lynn, D. H., and Keeling, P. J. (2014). Transcriptomics for Microbial Eukaryotes. *Curr. Biol.*, 24(22):R1081–R1082.
- Komer, B., Bergstra, J., and Eliasmith, C. (2014). Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn. *Proc. 13th Python Sci. Conf.*, (Scipy):33–39.
- Kopp, C., Domart-Coulon, I., Escrig, S., Humbel, B. M., Hignette, M., and Meibom, A. (2015). Subcellular investigation of photosynthesis-driven carbon and nitrogen assimilation and utilization in the symbiotic reef coral Pocillopora damicornis. *Submitted*, 6(1):1–9.
- Korfhage, C., Fricke, E., and Meier, A. (2015). Whole-Transcriptome Amplification of Single Cells for Next-Generation Sequencing. *Curr. Protoc. Mol. Biol.*, (July):7.20.1–7.20.19.
- Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., Soria, M., Wang, S., Akhunov, E., Uauy, C., and Dubcovsky, J. (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.*, 14:R66.
- Kreutz, M., Stoeck, T., and Foissner, W. (2012). Morphological and Molecular Characterization of Paramecium (Viridoparamecium nov. subgen.) chlorelligerum Kahl, 1935 (Ciliophora). *J. Eukaryot. Microbiol.*, 59(6).
- Krogh, a., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, 305(3):567–80.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.*, 142(2):573–603.
- Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.*, 84:283–289.
- Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. *Mol. Biol. Evol.*, 13(4):584–593.
- Küper, U., Meyer, C., Müller, V., Rachel, R., and Huber, H. (2010). Energized outer membrane and spatial separation of metabolic processes in the hyperthermophilic Archaeon Ignicoccus hospitalis. *Proc. Natl. Acad. Sci. U. S. A.*, 107:3152–3156.
- Laligne, C., Klotz, C., Garreau de Loubresse, N., Lemullois, M., Hori, M., Laurent, F. X., Papon, J. F., Louis, B., Cohen, J., and Koll, F. (2010). Bug22p, a Conserved Centrosomal/Ciliary Protein Also Present in Higher Plants, Is Required for an Effective Ciliary Stroke in Paramecium. *Eukaryot. Cell*, 9(4):645–655.

- Lalonde, S., Boles, E., Hellmann, H., Barker, L., Patrick, J., Frommer, W., and Ward, J. (1999). The dual function of sugar carriers. *Transport and sugar sensing*. *Plant Cell*, 11(April):707–726.
- LaMonte, G., Philip, N., Reardon, J., Lacsina, J., Majoros, W., Chapman, L., Thornburg, C., Telen, M., Ohler, U., Nicchitta, C., Haystead, T., and Chi, J.-T. (2012). Translocation of Sickle Cell Erythrocyte MicroRNAs into Plasmodium falciparum Inhibits Parasite Translation and Contributes to Malaria Resistance. *Cell Host Microbe*, 12:187–199.
- Lane, N. (2007). Mitochondria: Key to Complexity. In Martin, W. and Müller, M., editors, *Orig. Mitochondria Hydrog. SE - 2*, pages 13–38. Springer Berlin Heidelberg.
- Lange, M., Westermann, P., and Ahring, B. K. r. (2005). Archaea in protozoa and metazoa. *Appl. Microbiol. Biotechnol.*, 66:465–474.
- Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–9.
- Lapointe, F.-j., Kirsch, J. A. W., and Bleiwiess, R. (1994). Jackknifing of Weighted Trees: Validation of Phylogenies Reconstructed from Distance Matrices. *Mol. Phylogenet. Evol.*, 3(3):256–267.
- Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6):1095–1109.
- Lasken, R. S. (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.*, 10:510–516.
- Laskowski, R. a., Watson, J. D., and Thornton, J. M. (2005). ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res.*, 33:89–93.
- Lassmann, T., Hayashizaki, Y., and Daub, C. O. (2009). TagDust - A program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, 25(21):2839–2840.
- Lassmann, T. and Sonnhammer, E. L. L. (2005). Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(298):1–9.
- Le, H. S., Schulz, M. H., McCauley, B. M., Hinman, V. F., and Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic Acids Res.*, 41:1–11.
- Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7):1307–1320.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2323.
- Ledergerber, C. and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Brief. Bioinform.*, 12(5):489–497.

- Lee, E., Jung, H., Radivojac, P., Kim, J.-W., and Lee, D. (2009). Analysis of AML genes in dysregulated molecular networks. *BMC Bioinformatics*, 10 Suppl 9:S2.
- Legin, A. a., Schintlmeister, A., Jakupc, M. a., Galanski, M., Lichtscheidl, I., Wagner, M., and Keppler, B. K. (2014). NanoSIMS combined with fluorescence microscopy as a tool for subcellular imaging of isotopically labeled platinum-based anticancer drugs. *Chem. Sci.*, 5:3135.
- Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Res.*, 39(November 2010):D19–D21.
- Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., and De Clerck, O. (2012). Phylogeny and Molecular Evolution of the Green Algae. *CRC Crit. Rev. Plant Sci.*, 31(1):1–46.
- Leperc, G., Nowacki, M., Serrano, V., Gout, J.-F., Guglielmi, G., Duharcourt, S., and Meyer, E. (2009). Silencing-associated and meiosis-specific small RNA pathways in Paramecium tetraurelia. *Nucleic Acids Res.*, 37(3):903–915.
- Leung, H., Yiu, S. M., and Chin, F. (2014). IDBA-MTP: A Hybrid MetaTranscriptomic Assembler Based on Protein Information. In Sharan, R., editor, *Res. Comput. Mol. Biol. SE - 12*, volume 8394 of *Lecture Notes in Computer Science*, pages 160–172. Springer International Publishing.
- Leung, H. C. M., Yiu, S.-M., Parkinson, J., and Chin, F. Y. L. (2013). IDBA-MT: De Novo Assembler for Metatranscriptomic Data Generated from Next-Generation Sequencing Technology. *J. Comput. Biol.*, 20(7):540–550.
- Leung, T. L. F. and Poulin, R. (2008). Parasitism, commensalism, and mutualism: Exploring the many shades of symbioses. *Vie Milieu*, 58(2):107–115.
- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. a., Stewart, R., and Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.*, 15(553):1–21.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(January):btvo33–.
- Li, H. (2015). Seqtk.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Liang, H., Zen, K., Zhang, J., Zhang, C.-Y., and Chen, X. (2013). New roles for microRNAs in cross-species communication. *RNA Biol.*, 10(December):367–370.
- Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., and Bairoch, A. (2009). HAMAP: a database of completely sequenced

- microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.*, 37(October 2008):D471–8.
- Lineberger, R. D. (1980). Cryoprotection by glucose, sucrose, and raffinose to chloroplast thylakoids. *Plant Physiol.*, 65:298–304.
- Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., Thompson, J. F., Letovsky, S., Milos, P., and Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.*, 27(7):652–658.
- Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., Shi, Y., Zhang, H., Li, Y., Lam, T.-W., and Luo, R. (2012). COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, 28(22):2870–2874.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–137.
- Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C. S., Lignelli, E., Reinhardt, R., Höffner, S., Krüger, M., Tsionis, P. a., Borchardt, T., and Braun, T. (2013). A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biol.*, 14(2):R16.
- Lowe, E., Swalla, B., and Brown, C. T. (2014). Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species. *PeerJ Prepr.*, September:0–10.
- Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*, 9(10).
- Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605.
- MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.*, 5(January):1–7.
- Macmanes, M. D. (2015). Optimizing error correction of RNAseq reads. *bioRxiv Prepr.*, pages 1–4.
- Macmanes, M. D. and Eisen, M. B. (2013). Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ*, 1:e113.
- Maguire, F., Henriquez, F. L., Leonard, G., Dacks, J. B., Brown, M. W., and Richards, T. a. (2014). Complex patterns of gene fission in the eukaryotic folate biosynthesis pathway. *Genome Biol. Evol.*, 6(10):2709–20.
- Maguire, F. and Richards, T. (2014). Organelle Evolution: A Mosaic of ‘Mitochondrial’ Functions. *Curr. Biol.*, 24(11):R518–R520.
- Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

- Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. a., Volkmer, G. a., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome Sequencing in Microfabricated High-density Picolitre Reactors. *Nature*, 437(September):376–380.
- Margulis, L. (1998). *The Symbiotic Planet: A New Look At Evolution*. Basic Books, New York, 1st edition.
- Marker, S., Carradec, Q., Tanty, V., Arnaiz, O., and Meyer, E. (2014). A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia. *Nucleic Acids Res.*, 42(11):7268–7280.
- Marker, S., Le Mouël, A., Meyer, E., and Simon, M. (2010). Distinct RNA-dependent RNA polymerases are required for RNAi triggered by double-stranded RNA versus truncated transgenes in Paramecium tetraurelia. *Nucleic Acids Res.*, 38(12):4092–4107.
- Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads. *EMBnet J.*, 17(1):pp—-10.
- Martin, W. and Herrmann, R. G. (1998). Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? *Plant Physiol.*, 118(1):9–17.
- Martinez, J., Longdon, B., Bauer, S., Chan, Y.-S., Miller, W. J., Bourtzis, K., Teixeira, L., and Jiggins, F. M. (2014). Symbionts commonly provide broad spectrum resistance to viruses in insects: a comparative analysis of Wolbachia strains. *PLoS Pathog.*, In press(9).
- Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, 101(19):7287–7292.
- Matthew, L. (2004). RNAi for plant functional genomics. *Comp. Funct. Genomics*, 5:240–4.
- Mayer, K. M. and Forney, J. D. (1999). A Mutation in the Flanking 5'-TA-3' Dinucleotide Prevents Excision of an Internal Eliminated Sequence From the Paramecium tetraurelia Genome. *Genetics*, 151:597–604.
- McAuley, P. (1986). Uptake of Amino Acids by Cultured and Freshly Isolated Symbiotic Chlorella. *New Phytol.*, 104:415–427.
- McCUTCHAN, T. F., de la Cruz, V. F., Lal, A. A., Gunderson, J. H., Elwood, H. J., and Sogin, M. L. (1988). Primary sequences of two small subunit ribosomal RNA genes from Plasmodium falciparum. *Mol. Biochem. Parasitol.*, 28:63–68.

McFadden, G. I. (2014). Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb. Perspect. Biol.*, 6:ao16105.

McGrath, C. L., Gout, J.-F., Doak, T. G., Yanagi, A., and Lynch, M. (2014). Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. *Genetics*, 197(4):1417–28.

Mendes Maia, T., Gogendeau, D., Pennetier, C., Janke, C., and Basto, R. (2014). Bug22 influences cilium morphology and the post-translational modification of ciliary microtubules. *Biol. Open*, 3:138–151.

Meng, D., Cao, M., Oda, T., and Pan, J. (2014). The conserved ciliary protein Bug22 controls planar beating of *Chlamydomonas* flagella. *J. Cell Sci.*, 127:281–287.

Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. *Linux J.*, 239(March).

Milinkovitch, M. C., Leduc, G., Ada, J., Farnir, F., Georgest, M., and Hasegawa, M. (1996). Effects of Character Weighting and Species Sampling on Phylogeny Reconstruction: A Case Study Based on DNA Sequence Data in Cetaceans. *Genetics*, 144:1817–1833.

Milinkovitch, M. C. and Lyons-Weiler, J. (1998). Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. *Mol. Phylogen. Evol.*, 9(3):348–357.

Miller, R. G. (1974). The Jackknife - A Review.

Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. (2003). Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Syst. Biol.*, 52(5):674–683.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, a., and Punta, M. (2013). Challenges in homology search: HMMER₃ and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, 41(12):e121–e121.

Mitchell, T. E. (1997). *Machine Learning*. McGraw Hill.

Miwa, I. (2009). Regulation of Circadian Rhythms of *Paramecium bursaria* by Symbiotic Chlorella Species. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 4, pages 83–110. Springer.

Mochizuki, K. and Gorovsky, M. a. (2004). Conjugation-specific small RNAs in Tetrahymena have predicted properties of scan (scn) RNAs involved in genome rearrangement. *Genes Dev.*, 18(Nanney 1974):2068–2073.

Mockler, T. C. and Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85:1–15.

Moissl-Eichinger, C. and Huber, H. (2011). Archaeal symbionts and parasites. *Curr. Opin. Microbiol.*, 14:364–370.

Molnar, M. and Ilie, L. (2014). Correcting Illumina data. *Brief. Bioinform.*

- Monteith, K., Carroll, J. L., Seppi, K., and Martinez, T. (2011). Turning Bayesian model averaging into Bayesian model combination. *2011 Int. Jt. Conf. Neural Networks*, pages 2657–2663.
- Morf, L., Pearson, R. J., Wang, A. S., and Singh, U. (2013). Robust gene silencing mediated by antisense small RNAs in the pathogenic protist *Entamoeba histolytica*. *Nucleic Acids Res.*, 41(20):9424–37.
- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, a. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35:W182–W185.
- Muñoz Mérida, A., González-Plaza, J. J., Cañada, A., Blanco, A. M., García-López, M. D. C., Rodríguez, J. M., Pedrola, L., Sicardo, M. D., Hernández, M. L., De la Rosa, R., Belaj, A., Gil-Borja, M., Luque, F., Martínez-Rivas, J. M., Pisano, D. G., Trelles, O., Valpuesta, V., and Beuzón, C. R. (2013). De novo assembly and functional annotation of the olive (*Olea europaea*) transcriptome. *DNA Res.*, 20(January):93–108.
- Müller, T., Philippi, N., Dandekar, T., Schultz, J., and Wolf, M. (2007). Distinguishing species. *RNA*, 13:1469–1472.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.
- Murray, S. a., Patterson, D. J., and Thessen, A. E. (2012). Transcriptomics and microbial eukaryote diversity: A way forward. *Trends Ecol. Evol.*, 27(12):651–652.
- Muscatine, L. (1967). Glycerol Excretion by Symbiotic Algae from Corals and Tridacna and Its Control by the Host. *Science* (80-.), 156(April):516–519.
- Muscatine, L., Karakashian, S. J., and Karakashian, M. W. (1967). Soluble extracellular products of algae symbiotic with a ciliate, a sponge and a mutant hydra. *Comp. Biochem. Physiol.*, 20:1–12.
- Nabhan, A. R. and Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. *Brief. Bioinform.*, 13(1):122–134.
- Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, 24(98):34–35.
- Nakai, K. and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14:897–911.
- Nakajima, B. Y. K. and Nakaoka, Y. (1989). Circadian Change of Photosensitivity in Paramecium Bursaria. *J. Exp. Biol.*, 144:43–51.
- Nakao, M., Bono, H., Kawashima, S., Kamiya, T., Sato, K., Goto, S., and Kanehisa, M. (1999). Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG. *Genome Inform. Ser. Workshop Genome Inform.*, 10:94–103.
- Nakasugi, K., Crowhurst, R., Bally, J., and Waterhouse, P. (2014). Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant *Nicotiana benthamiana*. *PLoS One*, 9(3):e91776.

National Center for Biotechnology Information (2011). SRA Knowledge Base.

Nederbragt, A. J. (2013). Developments in Next-Generation Sequencing Technologies.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Nei, M. (1987). The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'. *Mol Biol Evol*, 4(4):406–425.

Neil, S. T. O. and Emrich, S. J. (2013). Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. *Stat. Decis. theory Relat. Top.*, pages 1–27.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Adv. Neural Inf. Process. Syst. 14 (NIPS 2001)*, pages 841–848. MIT Press.

Nielsen, H., Engelbrecht, J., Brunak, S., and Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, 10(1):1–6.

Niess, D., Reisser, W., and Wiessner, W. (1982a). Photobehaviour of Paramecium bursaria infected with different symbiotic and aposymbiotic species of Chlorella. *Planta*, 156:475–480.

Niess, D., Reisser, W., and Wiessner, W. (1982b). Photobehaviour of Paramecium bursaria infected with different symbiotic and aposymbiotic species of Chlorella. *Planta*, 156(5):475–480.

Nikolenko, S. I., Korobeynikov, A. I., and Alekseyev, M. a. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 14 Suppl 1(Suppl 1):S7.

Nishimoto, Y., Arisue, N., Kawai, S., Escalante, A. a., Horii, T., Tanabe, K., and Hashimoto, T. (2008). Evolution and phylogeny of the heterogeneous cytosolic SSU rRNA genes in the genus Plasmodium. *Mol. Phylogenet. Evol.*, 47:45–53.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, a., Shabalov, I., Smirnova, T., Grigoriev, I. V., and Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, 42(November 2013):D26–D31.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217.

Novakova, Lenka and Stepankova, O. (2006). Multidimensional clusters in RadViz. *Proc. 6th WSEAS Int. Conf. Simulation, Model. Optim.*, pages 470–475.

- Nowack, E. C. M., Vogel, H., Groth, M., Grossman, A. R., Melkonian, M., and Glöckner, G. (2011). Endosymbiotic gene transfer and transcriptional regulation of transferred genes in Paulinella chromatophora. *Mol. Biol. Evol.*, 28(1):407–422.
- Nowara, D., Gay, a., Lacomme, C., Shaw, J., Ridout, C., Douchkov, D., Hensel, G., Kumlehn, J., and Schweizer, P. (2010). HIGS: Host-Induced Gene Silencing in the Obligate Biotrophic Fungal Pathogen Blumeria graminis. *Plant Cell*, 22(September):3130–3141.
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. a., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotnik, A., Sirotnik, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., McLean, J. S., Lasken, R., Tesler, G., Alekseyev, M. a., and Pevzner, P. a. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.*, 20(10):714–37.
- O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2002). Streaming-Data Algorithms for High-Quality Clustering. In *Proc. 18th Int. Conf. Data Eng.*, page 0685. IEEE.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., Goto, S., Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M., and Goto, S. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 27(1):27–30.
- Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, 36(May):W423–W426.
- Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385.
- O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. In *Curr. Protoc. Mol. Biol.*, volume Unit 4.19, chapter 4. John Wiley & Sons, Inc.
- Ozsolak, F. and Milos, P. M. (2011). Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA*, 2(4):565–570.
- Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461(7265):814–818.
- O'Malley, M. a. (2015). Molecular organisms. *Biol. Philos.*
- Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X., Richter, D. J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., Li, C., Meyerson, M., and Sellers, W. R. (2004). Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.*, 32(9):e71.
- Pak, J. and Fire, A. (2007). Distinct Populations of Primary and Secondary Effectors During RNAi in *C. elegans*. *Science (80-.)*, 315:241–244.

- Palmer, L. E., Dejori, M., Bolanos, R., and Fasulo, D. (2010). Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction. *BMC Bioinformatics*, 11:33.
- Parker, R. C. (1926). Symbiosis in Paramecium bursaria. *J. Exp. Zool.*, 46(1):1–12.
- Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.*, 33:1065–1076.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.*, 12:2825–2830.
- Pedrioli, P. G. a., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, 22(11):1459–1466.
- Pell, J., Hintze, a., Canino-Koning, R., Howe, a., Tiedje, J. M., and Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci.*, 109(33):13272–13277.
- Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, 12(8):780–786.
- Peng, Y., Leung, H., Yiu, S. M., and Chin, F. (2010). IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In Berger, B., editor, *Res. Comput. Mol. Biol. SE - 28*, volume 6044 of *Lecture Notes in Computer Science*, pages 426–440. Springer Berlin Heidelberg.
- Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–8.
- Peng, Y., Leung, H. C. M., Yiu, S.-M., Lv, M.-J., Zhu, X.-G., and Chin, F. Y. L. (2013). IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29:1326–1334.
- Peralta, H., Guerrero, G., Aguilar, A., and Mora, J. (2011). Sequence variability of Rhizobiales orthologs and relationship with physico-chemical characteristics of proteins. *Biol. Direct*, 6(1):48.
- Pérez, M. T., Dolan, J. R., and Fukai, E. (1997). Planktonic oligotrich ciliates in the NW Mediterranean: growth rates and consumption by copepods. *Mar. Ecol. Prog. Ser.*, 155:89–101.
- Perez-Cobas, a. E., Gosalbes, M. J., Friedrichs, a., Knecht, H., Artacho, a., Eismann, K., Otto, W., Rojo, D., Bargiela, R., von Bergen, M., Neulinger, S. C., Daumer, C., Heinsen, F.-a., Latorre, a., Barbas, C., Seifert, J., dos Santos,

- V. M., Ott, S. J., Ferrer, M., and Moya, a. (2013). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, 62:1591–1601.
- Peters, L. and Meister, G. (2007). Argonaute Proteins: Mediators of RNA Silencing. *Mol. Cell*, 26:611–623.
- Petersen, T. N., Brunak, S. r., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, 8(10):785–6.
- Phadke, S. S. and Zufall, R. a. (2009). Rapid diversification of mating systems in ciliates. *Biol. J. Linn. Soc.*, 98:187–197.
- Pillmann, a., Woolcott, G. W., Olsen, J. L., Stam, W. T., and King, R. J. (1997). Inter- and intraspecific genetic variation in Caulerpa (Chlorophyta) based on nuclear rDNA ITS sequences. *Eur. J. Phycol.*, 32(December):379–386.
- Pop, M. (2009). Genome assembly reborn: Recent computational challenges. *Brief. Bioinform.*, 10(4):354–366.
- Poretsky, R. S., Bano, N., Buchan, A., Kleikemper, J., Pickering, M., Pate, W. M., Moran, M. A., Hollibaugh, J. T., and Lecleir, G. (2005). Analysis of Microbial Gene Transcripts in Environmental Samples. *Appl. Environ. Microbiol.*, 71(7):4121–4126.
- Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. *Mol. Biol. Evol.*, 25:1253–1256.
- Pound, R. (1893). Symbiosis and Mutualism. *Am. Nat.*, 27:509.
- Prell, J. and Poole, P. (2006). Metabolic changes of rhizobia in legume nodules. *Trends Microbiol.*, 14(4):161–168.
- Prescott, D. M. (1994). The DNA of ciliated protozoa. *Microbiol. Rev.*, 58(2):233–267.
- Preston, C. M., Wu, K. Y., Molinski, T. F., and DeLong, E. F. (1996). A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc. Natl. Acad. Sci. U. S. A.*, 93(June):6241–6246.
- Price, D. C., Chan, C. X., Yoon, H. S., Yang, E. C., Qiu, H., Weber, A. P. M., Schwacke, R., Gross, J., Blouin, N. a., and Lane, C. (2012). Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science* (80-.), 335(March):843–847.
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490.
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35:501–504.
- Putt, M. (1990). Abundance, chlorophyll content and photosynthetic rates of ciliates in the Nordic Seas during summer. *Deep. Res.*, 37(11):1713–1731.

- Quail, M. a., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1):341.
- Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. In *Curr. Protoc. Bioinforma.* John Wiley & Sons, Inc.
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.
- Quinlan, J. R. (1986). Induction of Decision Trees. *Mach. Learn.*, 1(1):81–106.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Raikov, I. (1995). Structure and Genetic Organization of the Polyploid Macronucleus of Ciliates: a Comparative Review. *Acta Protozool.*, 34:151–171.
- Raj, A. and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226.
- Rambaut, A. and Drummond, A. J. (2007). Tracer v1. 4.
- Raven, J. (1997). Phagotrophy in phototrophs. *Limnol. Ocean.*, 42(1):198–205.
- Redelings, B. D. and Suchard, M. a. (2005). Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3):401–418.
- Regalado, A. (2014). EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. *MIT Technol. Rev.*, (The Year in Review: Health Care).
- Reisser, W. (1980). The Metabolic Interactions Between Paramecium bursaria Ehrbg. and Chlorella spec. in the Paramecium bursaria-Symbiosis. *Arch. Microbiol.*, 125:291–293.
- Reisser, W., Burbank, D. E., Meints, S. M., Meints, R. H., Becker, B., and Van Etten, J. L. (1988). A comparison of viruses infecting two different Chlorella-like green algae. *Virology*, 167(1):143–149.
- Richards, T. a., Soanes, D. M., Foster, P. G., Leonard, G., Thornton, C. R., and Talbot, N. J. (2009). Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. *Plant Cell*, 21(July):1897–1911.
- Ris, H. and Plaut, W. (1962). Ultrastructure of DNA-containing areas in the chloroplast of Chlamydomonas. *J. Cell Biol.*, 13:383–391.
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K.,

- She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. a., Hirst, M., Marra, M. a., Jones, S. J. M., Hoodless, P. a., and Birol, I. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods*, 7(11):909–912.
- Rockwell, N. C., Lagarias, J. C., and Bhattacharya, D. (2014). Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Front. Ecol. Evol.*, 2(October):1–13.
- Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, a., Hohna, S., Larget, B., Liu, L., Suchard, M. a., and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.*, 61(3):539–542.
- Rosenblatt, M. (1956). Remarks on Some Non-parametric Estimates of a Density Function. *Ann. Math. Stat.*, 27(3):832–837.
- Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. *Nucleic Acids Res.*, 32:W321–W326.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.
- Rowe, J. M., Jeanniard, A., Gurnon, J. R., Xia, Y., Dunigan, D. D., Van Etten, J. L., and Blanc, G. (2014). Global analysis of Chlorella variabilis NC64A mRNA profiles during the early phase of Paramecium bursaria chlorella virus-1 infection. *PLoS One*, 9(3).
- Rusch, S. L. and Kendall, D. A. (1995). Protein transport via amino-terminal targeting sequences: common themes in diverse systems (Review). *Mol. Membr. Biol.*, 12(4):295–307.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge 2010. *arXiv:1409.0575*.
- Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behav. Ecol.*, 17(May):688–690.
- Saier, M. H., Reddy, V. S., Tamang, D. G., and Västermark, A. k. (2014). The transporter classification database. *Nucleic Acids Res.*, 42(November 2013):251–258.
- Saier, M. H., Tran, C. V., and Barabote, R. D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.*, 34:D181–D186.
- Saier, M. H., Yen, M. R., Noto, K., Tamang, D. G., and Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic Acids Res.*, 37(November 2008):D274–D278.
- Saitou, N. and Nei, M. (1987). The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'. *Mol Biol Evol*, 4(4):406–425.

- Saji, M. and Oosawa, F. (1974). Mechanism of Photoaccumulation in Paramecium bursaria. *J. Protozool.*, 22(4):57–65.
- Salim, H. M. W., Ring, K. L., and Cavalcanti, a. R. O. (2008). Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: Tetrahymena thermophila and Paramecium tetraurelia. *Protist*, 159(April):283–298.
- Sana, T. R., Roark, J. C., Li, X., Waddell, K., and Fischer, S. M. (2008). Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS. *J. Biomol. Tech.*, 19:258–266.
- Sanderson, C. (2010). Conrad Sanderson. Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. *NICTA*, Technical.
- Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, a. R., Fiddes, C. a., Hutchison, C. a., Slocombe, P. M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265:687–695.
- Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94:441–448.
- Sanger, F., Nicklen, S., and Coulson, a. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467.
- Sassera, D., Beninati, T., Bandi, C., Bouman, E. a. P., Sacchi, L., Fabbri, M., and Lo, N. (2006). Candidatus Midichloria mitochondrii, an endosymbiont of the Ixodes ricinus with a unique intramitochondrial lifestyle. *Int. J. Syst. Evol. Microbiol.*, 56:2535–2540.
- Sato, S. (2011). The apicomplexan plastid and its evolution. *Cell. Mol. Life Sci.*, 68:1285–1296.
- Schatz, G. and Dobberstein, B. (1996). Common Principles of Protein Translocation Across Membranes. *Science* (80-.), 271(5255):1519–1526.
- Schatz, M. C., Delcher, a. L., Roberts, M., Marcqais, G., Pop, M., and Yorke, J. a. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms Steven L. Salzberg, Adam M. Phillippe, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen. *Genome Res.*, 22:557–567.
- Schmieder, R. and Edwards, R. (2011). Quality Control and Preprocessing of Metagenomic Datasets. *Bioinformatics*, 27(6):863–864.
- Schrallhammer, M. and Schweikert, M. (2009). The Killer Effect of Paramecium and Its Causative Agents. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 9, pages 227–246. Springer.
- Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.*, 95(May):5857–5864.

- Schultz, J. and Wolf, M. (2009). ITS₂ sequence–structure analysis in phylogenetics: A how-to manual for molecular systematics. *Mol. Phylogenet. Evol.*, 52(2):520–523.
- Schulz, F. and Horn, M. (2015). Intranuclear bacteria: inside the cellular control center of eukaryotes. *Trends Cell Biol.*, pages 1–8.
- Schulz, F., Lagkouvardos, I., Wascher, F., Aistleitner, K., Kostanjšek, R., and Horn, M. (2014). Life in an unusual intracellular niche: a bacterial symbiont infecting the nucleus of amoebae. *ISME J.*, pages 1634–1644.
- Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–92.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6:461–464.
- Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.
- Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, 3(3):246–251.
- Setter, T. L. and Greenway, H. (1979). Growth and osmoregulation of Chlorella emersonii in NaCl and neutral osmotica. *Plant Physiol.*, 6:47–60.
- Shabalina, S. A. and Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. *Trends Ecol. Evol.*, 23(10):578–587.
- Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. J., Gennert, D., Gnrke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–40.
- Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145.
- Shi, D.-Q. (2005). SLOW WALKER1, Essential for Gametogenesis in Arabidopsis, Encodes a WD40 Protein Involved in 18S Ribosomal RNA Biogenesis. *Plant Cell Online*, 17(August):2340–2354.
- Shihira, I. and Krauss, R. W. (1965). *Chlorella: physiology and taxonomy of forty-one isolates*. University of Maryland.
- Siegel, R. W. (1963). New result on genetics of mating types in Paramecium bursaria. *Genet. Res.*, 4:132–142.
- Siegel, R. W. and Karakashian, S. J. (1959). Dissociation and restoration of endocellular symbiosis in Paramecium-Bursaria. In *Anat. Rec.*, volume 134, page 639. WILEY-LISS DIV JOHN WILEY & SONS INC, 605 THIRD AVE, NEW YORK, NY 10158-0012.
- Siegel, R. W. and Larison, L. L. (1960). The Genic Control of Mating Types in Paramecium Bursaria. *Proc. Natl. Acad. Sci. U. S. A.*, 46(variety 8):344–349.

- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7(539).
- Sijen, T., Steiner, F. A., Thijssen, K. L., and Plasterk, R. H. A. (2007). Secondary siRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class. *Science* (80-.), 315:244–247.
- Simmons, M. P. (2003). How Meaningful Are Bayesian Support Values? *Mol. Biol. Evol.*, 21(1):188–199.
- Simpson, E. (1951). The Interpretation of Interaction in Contingency Tables. *J. R. Stat. Soc. Ser. B*, 13(2):238–241.
- Sims, D., Ilott, N. E., Sansom, S. N., Sudbery, I. M., Johnson, J. S., Fawcett, K. a., Berlanga-Taylor, A. J., Luna-Valero, S., Ponting, C. P., and Heger, A. (2014). CGAT: Computational genomics analysis toolkit. *Bioinformatics*, 30(9):1290–1291.
- Smith, C., Elizabeth, J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using Nonlinear Peak Alignment, Matching, and Identification. *ACS Publ.*, 78(3):779–787.
- Smith, C. A., Maille, G. O., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005a). METLIN: A Metabolite Mass Spectral Database. *Ther. Drug Monit.*, 27(6).
- Smith, J. C., Northey, J. G. B., Garg, J., Pearlman, R. E., and Siu, K. W. M. (2005b). Robust method for proteome analysis by MS/MS using an entire translated genome: Demonstration on the ciliome of Tetrahymena thermophila. *J. Proteome Res.*, 4:909–919.
- Smith, L., Sanders, J., Kaiser, R., Hughes, P., Dodd, C., Connell, C., Heiner, C., Kent, S., and Hood, L. (1986). Fluorescence Detection in Automated DNA Sequence Analysis. *Nature*, 321(June):674–679.
- Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J., and Hood, L. E. (1985). The Synthesis of Oligonucleotides Containing an Alipathic Amino Group at the 5' Terminus: Synthesis of Fluorescent DNA Primers for use in DNA Sequence Analysis. *Nucleic Acids Res.*, 13(7):2399–2412.
- Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.
- Smith-unna, R., Boursnell, C., Patro, R., Hibberd, J. M., and Kelly, S. (2015). TransRate: reference free quality assessment of de-novo transcriptome assemblies. *bioRxiv Prepr*, June:1–25.
- Sommaruga, R. and Sonntag, B. (2009). Photobiological Aspects of the Mutualistic Association Between Paramecium bursaria and Chlorella. In Fujishima, M., editor, *Endosymbionts in Paramecium*, volume 12 of *Microbiology Monographs*, chapter 5, pages 111–130. Springer.
- Sonneborn, T. (1950). Methods in the general biology and genetics of Paramecium aurelia. *J. Exp. Zool.*, 113(1):87–147.

- Sonneborn, T. M. (1970). Methods in Paramecium Research. In *Methods Cell Biol.*, volume 4, chapter 12, pages 241–339.
- Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Sixth Int. Conf. Intell. Syst. Mol. Biol.*, 6:175–82.
- Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.*, 1(4):1965–1970.
- Stamatakis, a. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.
- Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst. Biol.*, 57(5):758–771.
- Stern, R. F., Andersen, R. a., Jameson, I., Küpper, F. C., Coffroth, M.-A., Vaulot, D., Le Gall, F., Véron, B., Brand, J. J., Skelton, H., Kasai, F., Lilly, E. L., and Keeling, P. J. (2012). Evaluating the ribosomal internal transcribed spacer (ITS) as a candidate dinoflagellate barcode marker. *PLoS One*, 7(8):e42780.
- Stiller, J. W., Schreiber, J., Yue, J., Guo, H., Ding, Q., and Huang, J. (2014). The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.*, 5:1–7.
- Stocking, C. and Gifford Jr., E. (1959). Incorporation of thymidine into chloroplasts of Spirogyra. *Biochem. Biophys. Res. Commun.*, 1(3):159–164.
- Stöver, B. C. and Müller, K. F. (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11:7.
- Suchard, M. a. and Redelings, B. D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16):2047–8.
- Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Commun. Stat. Methods*, 7(1):13–26.
- Sullivan, J. and Joyce, P. (2005). Model Selection in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, 36(May):445–466.
- Summerer, M., Sonntag, B., Hörtnagl, P., and Sommaruga, R. (2009). Symbiotic ciliates receive protection against UV damage from their algae: a test with Paramecium bursaria and Chlorella. *Protist*, 160(2):233–43.
- Summerer, M., Sonntag, B., and Sommaruga, R. (2007). An experimental test of the symbiosis specificity between the ciliate Paramecium bursaria and strains of the unicellular green alga Chlorella. *Environ. Microbiol.*, 9(8):2117–22.
- Summerer, M., Sonntag, B., and Sommaruga, R. (2008). CILIATE-SYMBIONT SPECIFICITY OF FRESHWATER ENDOSYMBIOTIC <i>CHLORELLA</i> (TREBOUXIOPHYCEAE, CHLOROPHYTA). *J. Phycol.*, 44:77–84.

- Sung, W., Tucker, A. E., Doak, T. G., Choi, E., Thomas, W. K., and Lynch, M. (2012). Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proc. Natl. Acad. Sci. U. S. A.*, 109(47):19339–44.
- Suzuki, S., Kakuta, M., Ishida, T., and Akiyama, Y. (2014). GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One*, 9(8):e103833.
- Suzuki, S., Kakuta, M., Ishida, T., and Akiyama, Y. (2015). Faster sequence homology searches by clustering subsequences. *Bioinformatics*, 31(November 2014):1183–1190.
- Swart, E. C., Bracht, J. R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J. S., Goldman, A. D., Nowacki, M., Schotanus, K., Jung, S., Fulton, R. S., Ly, A., McGrath, S., Haub, K., Wiggins, J. L., Storto, D., Matese, J. C., Parsons, L., Chang, W. J., Bowen, M. S., Stover, N. a., Jones, T. a., Eddy, S. R., Herrick, G. a., Doak, T. G., Wilson, R. K., Mardis, E. R., and Landweber, L. F. (2013). The *Oxytricha trifallax* Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. *PLoS Biol.*, 11(1).
- Takahashi, T., Shirai, Y., Kosaka, T., and Hosoya, H. (2007). Arrest of cytoplasmic streaming induces algal proliferation in green paramecia. *PLoS One*, 2(12):e1352.
- Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, 56(4):564–577.
- Tanaka, M., Murata-Hori, M., Kadono, T., Yamada, T., Kawano, T., Kosaka, T., and Hosoya, H. (2002). Complete elimination of endosymbiotic algae from *Paramecium bursaria* and its confirmation by diagnostic PCR. *Acta Protozool.*, 41:255–261.
- Tange, O. (2011). GNU Parallel - The Command-Line Power Tool. ;login USENIX Mag., 36(1):42–47.
- Tanner, W., Haass, D., Decker, M., Loos, E., Komor, B., and Komor, E. (1974). Active Hexose Transport in *Chlorella vulgaris*. In Zimmermann, U. and Dainty, J., editors, *Membr. Transp. Plants SE - 28*, pages 202–208. Springer Berlin Heidelberg.
- Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., and Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.*, 30(1):27–30.
- Tautenhahn, R., Bottcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9:504.
- Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., and Siuzdak, G. (2012a). An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol*, 30(9):826–828.
- Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012b). XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.*, 84:5035–9.
- Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences.

- Thomas, P. D. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.*, 31(1):334–341.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680.
- Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PLoS One*, 6(3):e18093.
- Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124.
- Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., Schöck, U., Guthke, R., and Kuchler, K. (2012). An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front. Microbiol.*, 3(March):1–14.
- Timmis, J. N., Ayliffe, M. a., Huang, C. Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.*, 5(February):123–135.
- Timmons, L., Court, D. L., and Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene*, 263:103–112.
- Tindall, K. and Kunkel, T. (1988). Fidelity of DNA Synthesis by the *Thermus aquaticus* DNA Polymerase. *Biochemistry*, 27:6008–6013.
- Tonooka, Y. and Watanabe, T. (2002). A natural strain of *Paramecium bursaria* lacking symbiotic algae. *Eur. J. Protistol.*, 38:55–58.
- Torgerson, W. (1952). Multidimensional Scaling: I Theory and Method. *Psychometrika*, 17(4):401–419.
- Torres-Machorro, A. L., Hernández, R., Cevallos, A. M., and López-Villaseñor, I. (2010). Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny? *FEMS Microbiol. Rev.*, 34:59–86.
- Trapnell, C., Williams, B. a., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2011). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat. Biotechnol.*, 28(5):511–515.
- Tsiatis, A. C., Norris-Kirby, A., Rich, R. G., Hafez, M. J., Gocke, C. D., Eshleman, J. R., and Murphy, K. M. (2010). Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. *J. Mol. Diagn.*, 12(4):425–432.
- Valentine, M. S., Rajendran, A., Yano, J., Weeraratne, S. D., Beisson, J., Cohen, J., Koll, F., and Van Houten, J. (2012). Paramecium BBS genes are key to presence of channels in Cilia. *Cilia*, 1(1):16.

- Van Etten, J. L., Burbank, D. E., KuczmarSKI, D., and Meints, R. H. (1983). Virus Infection of Culturable Chlorella-Like Algae and Development of a Plaque Assay. *Science* (80-.), 219:994–996.
- van Gorkom, H. J. (1974). Identification of the reduced primary electron acceptor of Photosystem II as a bound semiquinone anion. *Biochim. Biophys. Acta - Bioenerg.*, 347:439–442.
- Vapnik, V. and Lerner, a. (1963). Pattern recognition using generalized portrait method. *Autom. Remote Control*, 24:774–780.
- Vapnik, V. N. and Kotz, S. (1982). *Estimation of dependences based on empirical data*, volume 41. Springer-Verlag New York.
- Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18 Suppl 1:S276–S284.
- Vijay, N., Poelstra, J. W., Künstner, A., and Wolf, J. B. W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, 22:620–634.
- Vitter, J. S. (1985). Random Sampling with a Reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57.
- Vogt, G. (1992). Enclosure of bacteria by the rough endoplasmic reticulum of shrimp hepatopancreas cells. *Protoplasma*, 171:89–96.
- von Dohlen, C. D., Kohler, S., Alsop, S. T., and McManus, W. R. (2001). Mealybug β -proteobacterial endosymbionts contain γ -proteobacterial symbionts. *Nature*, 412(July):433–436.
- von Heijne, G. (2006). Membrane-protein topology. *Nat. Rev. Mol. Cell Biol.*, 7(December):909–918.
- Vorobyev, K., Andronov, E., Skoblo, I., Migunova, A., and Kvitko, K. (2009). An atypical Chlorella symbiont from Paramecium bursaria. *Protistology*, 6(1):39–44.
- Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.*, 1:337–348.
- Wang, X. V., Blades, N., Ding, J., Sultana, R., and Parmigiani, G. (2012). Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, 13:185.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10:57–63.
- Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., Miles, A., Augspurger, T., Yarkoni, T., Evans, C., Wehner, D., Rocher, L., Megies, T., Coelho, L. P., Ziegler, E., Hoppe, T., Seabold, S., Pascual, S., Cloud, P., Koskinen, M., Hausler, C., Kjemmett, Milajevs, D., Qalieh, A., Allan, D., and Meyer, K. (2015). seaborn: v0.6.0 (June 2015).
- Wass, M. N., Barton, G., and Sternberg, M. J. E. (2012). CombFunc: predicting protein function using heterogeneous data sources. *Nucleic Acids Res.*, 40(May):W466–70.

- Waterhouse, P. M., Wang, M. B., and Lough, T. (2001). Gene silencing as an adaptive defence against viruses. *Nature*, 411:834–42.
- Watnick, P. and Kolter, R. (2000). Biofilm, city of microbes. *J. Bacteriol.*, 182(10):2675–2679.
- Wehrens, R., Weingart, G., and Mattivi, F. (2014). metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *J. Chromatogr. B*, 966:109–116.
- Weiberg, A., Bellinger, M., and Jin, H. (2015). Conversations between kingdoms: small RNAs. *Curr. Opin. Biotechnol.*, 32C:207–215.
- Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H.-D., and Jin, H. (2013). Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science*, 342(October):118–23.
- Weis, D. S. (1984). The Effect of Accumulated Time of Separate Cultivation on the Frequency of Infection of Aposymbiotic Ciliates by Symbiotic Algae in *Paramebium bursaria*. *J. Protozool.*, 31(4):A13—A14.
- Weis, D. S. and Ayala, A. (1976). Effect of Exposure Period and Algal Concentration on the Frequency of Infection of Aposymbiotic Ciliates by Symbiotic Algae from *Paramecium bursaria*. *J. Protozool.*, 26(2):245–248.
- Welch, B. (1947). The Generalisation of the 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35.
- Wernegreen, J. J. (2012). Endosymbiosis. *Curr. Biol.*, 22:555–561.
- Westermann, A. J., Gorski, S. a., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.*, 10(9):618–630.
- Whangbo, J. S. and Hunter, C. P. (2008). Environmental RNA interference. *Trends Genet.*, 24(April):297–305.
- White, T. J., Bruns, T., Lee, S., Taylor, J. W., and Others (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protoc. a Guid. to methods Appl.*, 18:315–322.
- Wichterman, R. (1946). TIME RELATIONSHIPS OF THE NUCLEAR BEHAVIOR IN THE CONJUGATION OF GREEN AND COLORLESS PARAMECIUM-BURSARIA. In *Anat. Rec.*, volume 94, pages 381–382. WILEY-LISS DIV JOHN WILEY & SONS INC, 605 THIRD AVE, NEW YORK, NY 10158-0012.
- Wichterman, R. (1948). The Biological Effects of X-Rays on Mating Types and Conjugation of Paramecium Bursaria. *Biol. Bull.*, 94(2):113–127.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.
- Wickham, H. and Francois, R. (2014). *dplyr: A Grammar of Data Manipulation*.
- Wilcox, L. W. (1986). Prokaryotic endosymbionts in the chloroplast stroma of the dinoflagellate *Woloszynskia pascheri*. *Protoplasma*, 135:71–79.

- Wilhelm, B. T. and Landry, J.-R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257.
- Williams, T., Kelley, C., and others, M. (2010). Gnuplot 4.4: an interactive plotting program. <http://gnuplot.sourceforge.net/>.
- Winnebenninkx, B., Backeljau, T., and De Wachter, R. (1993). Extraction of high molecular weight DNA from molluscs. *Trends Genet.*, 9(127):407.
- Wolpert, D. and Macready, W. G. (1995). No free lunch theorems for search. *Tech. Rep. SFI-TR-95-02-010*, pages 1–38.
- Wolpert, D. H. (1996). The Existence of A Priori Distinctions Between Learning Algorithms. *Neural Comput.*, 8(7):1391–1420.
- Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1):67–82.
- Wrede, C., Dreier, A., Kokoschka, S., and Hoppert, M. (2012). Archaea in symbioses. *Archaea*, 2012:596846.
- Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., and Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods*, 11(1):41–6.
- Wu, L. and Belasco, J. G. (2008). Let Me Count the Ways: Mechanisms of Gene Regulation by miRNAs and siRNAs. *Mol. Cell*, 29:1–7.
- Wu, R., Shengen, Y., Shan, Y., Dang, Q., and Sun, G. (2015). Deep Image: Scaling up Image Recognition. *arXiv:1501.02876*.
- Xiang, T., Nelson, W., Rodriguez, J., Tolleter, D., and Grossman, A. R. (2015). Symbiodinium transcriptome and global responses of cells to immediate changes in light intensity when grown under autotrophic or mixotrophic conditions. *Plant J.*, 82:67–80.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. K. S., and Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666.
- Yanagi, A. (2004). Autogamy is induced in Paramecium bursaria by methyl cellulose. *Eur. J. Protistol.*, 40:313–315.
- Yang, X., Chockalingam, S. P., and Aluru, S. (2013a). A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform.*, 14(1):56–66.
- Yang, X., Chockalingam, S. P., and Aluru, S. (2013b). A survey of error-correction methods for next-generation sequencing. *Brief. Bioinform.*, 14(1):56–66.

- Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10(6):1396–1401.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–314.
- Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *TREE*, 11(9):367–372.
- Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, 13(May):303–314.
- Yao, M.-C. and Chao, J.-L. (2005). RNA-guided DNA deletion in Tetrahymena: an RNAi-based mechanism for programmed genome rearrangements. *Annu. Rev. Genet.*, 39:537–59.
- Yashchenko, V. V., Gavrilova, O. V., Rautian, M. S., and Jakobsen, K. S. (2012). Association of Paramecium bursaria Chlorella viruses with Paramecium bursaria cells: ultrastructural studies. *Eur. J. Protistol.*, 48(2):149–59.
- Yellowlees, D., Rees, T. A. V., and Leggat, W. (2008). Metabolic interactions between algal symbionts and invertebrate hosts. *Plant. Cell Environ.*, 31:679–694.
- Yule, G. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134.
- Zdobnov, E. M. and Apweiler, R. (2001). Signature-recognition methods in InterPro. 17(9):847–848.
- Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18:821–9.
- Zhang, Q., Awad, S., and Brown, C. (2015). Crossing the streams : a framework for streaming analysis of short DNA sequencing reads PrePrints analysis of short DNA sequencing reads PrePrints. pages 0–27.
- Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C., and Brown, C. T. (2014). These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure. *PLoS One*, 9(7):e101271.
- Zhang, T., Ramakrishnan, R., and Livny, M. (1996). BIRCH: An Efficient Data Clustering Databases Method for Very Large. *ACM SIGMOD Int. Conf. Manag. Data*, 1:103–114.
- Zhou, C., Mao, F., Yin, Y., Huang, J., Gogarten, J. P., and Xu, Y. (2014). AST: An automated sequence-sampling method for improving the taxonomic diversity of gene phylogenetic trees. *PLoS One*, 9(6):2–10.
- Zhou, S., Wei, F., Nguyen, J., Bechner, M., Potamousis, K., Goldstein, S., Pape, L., Mehan, M. R., Churas, C., Pasternak, S., Forrest, D. K., Wise, R., Ware, D., Wing, R. a., Waterman, M. S., Livny, M., and Schwartz, D. C. (2009). A Single Molecule Scaffold for the Maize Genome. *PLoS Genet.*, 5(11):e1000711.
- Zhou, Y., Brinkmann, H., Rodrigue, N., Lartillot, N., and Philippe, H. (2010). A dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Mol. Biol. Evol.*, 27(2):371–384.
- Zwickl, D. J. and Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.*, 51(4):588–598.

Appendices

A

Appendix 1

A.1 ARBORETUM CLASSIFIER COMPARISON

A.1.1 GENOMES USED

Genomes used in transcript binning pipeline. Genomes were chosen to be a representative of the sampled diversity of the eukaryotic tree of life as possible:

- *Arabidopsis thaliana*
- *Chlamydomonas reinhardtii*
- *Ostreococcus tauri*
- *Micromonas pusilla CCMP1545*
- *Chlorella variabilis NC64A*
- *Chlorella vulgaris C-169*
- *Physcomitrella patens*
- *Saccharomyces cerevisiae S288C*
- *Neurospora crassa OR74A*

- *Homo sapiens*
- *Mus musculus*
- *Dictyostelium discoideum*
- *Paramecium caudatum*
- *Paramecium tetraurelia*
- *Tetrahymena thermophila macronucleus*
- *Oxytricha trifallax*
- *Toxoplasma gondii*
- *Guillardia theta*
- *Bigelowiella natans*
- *Emiliania huxleyi CCMP1516*
- *Aureococcus anophagefferens*
- *Ectocarpus siliculosus*
- *Schizosaccharomyces pombe*
- *Bacillus cereus ATCC 14579*
- *Escherichia coli str. K-12 substr. MG1655*
- *Escherichia coli O157 H7 str. Sakai*
- *Salmonella enterica subsp. enterica serovar Typhi str. CT18*
- *Amycolatopsis mediterranei U32*
- *Aquifex aeolicus VF5*
- *Borrelia burgdorferi B31*
- *Chlamydophila pneumoniae CWL029*
- *Chlorobium tepidum TLS*
- *Deinococcus radiodurans R2*
- *Caulobacter crescentus CB15*
- *Sulfolobus islandicus M.14.25*
- *Nanoarchaeum equitans Kin4-M*

- *Haloferax mediterranei* ATCC 33500
- *Methanococcus maripaludis* S2
- *Cenarchaeum symbiosum* A

A.1.2 CLASSIFICATION REPORTS

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.96 | 0.84 | 0.90 | 156 |
| food | 0.98 | 0.99 | 0.99 | 426 |
| host | 0.90 | 0.99 | 0.94 | 787 |
| endosymbiont | 0.97 | 0.99 | 0.98 | 359 |
| avg / total | 0.96 | 0.96 | 0.95 | 1728 |

Table A.1.1: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_neighbors=50, p=2, weights='uniform')

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.94 | 0.85 | 0.90 | 156 |
| food | 0.98 | 1.00 | 0.99 | 426 |
| host | 0.92 | 0.96 | 0.94 | 787 |
| endosymbiont | 0.96 | 0.99 | 0.98 | 359 |
| avg / total | 0.95 | 0.95 | 0.95 | 1728 |

Table A.1.2: LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.0001, verbose=0)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.43 | 0.94 | 0.59 | 156 |
| food | 0.93 | 0.12 | 0.22 | 426 |
| host | 0.95 | 0.85 | 0.90 | 787 |
| endosymbiont | 0.99 | 0.97 | 0.98 | 359 |
| avg / total | 0.83 | 0.70 | 0.65 | 1728 |

Table A.1.3: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, degree=3, gamma=0.0, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.90 | 0.83 | 0.86 | 156 |
| food | 0.97 | 0.98 | 0.98 | 426 |
| host | 0.89 | 0.93 | 0.91 | 787 |
| endosymbiont | 0.98 | 0.99 | 0.99 | 359 |
| avg / total | 0.94 | 0.94 | 0.94 | 1728 |

Table A.1.4: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, random_state=None, splitter='best')

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.84 | 0.83 | 0.83 | 156 |
| food | 0.97 | 0.98 | 0.97 | 426 |
| host | 0.87 | 0.93 | 0.88 | 787 |
| endosymbiont | 0.97 | 0.97 | 0.97 | 359 |
| avg / total | 0.92 | 0.92 | 0.92 | 1728 |

Table A.1.5: ExtraTreeClassifier(class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, random_state=None, splitter='random')

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.89 | 0.91 | 0.90 | 156 |
| food | 0.98 | 1.00 | 0.99 | 426 |
| host | 0.95 | 0.92 | 0.94 | 787 |
| endosymbiont | 0.97 | 0.96 | 0.97 | 359 |
| avg / total | 0.95 | 0.95 | 0.95 | 1728 |

Table A.1.6: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.90 | 0.83 | 0.86 | 156 |
| food | 0.98 | 1.00 | 0.99 | 426 |
| host | 0.88 | 0.93 | 0.90 | 787 |
| endosymbiont | 0.98 | 0.97 | 0.98 | 359 |
| avg / total | 0.94 | 0.94 | 0.94 | 1728 |

Table A.1.7: AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.31 | 0.99 | 0.47 | 156 |
| food | 0.99 | 0.79 | 0.88 | 426 |
| host | 1.00 | 0.04 | 0.08 | 787 |
| endosymbiont | 0.00 | 0.00 | 0.00 | 359 |
| avg / total | 0.58 | 0.47 | 0.38 | 1728 |

Table A.1.8: LDA(n_components=None, priors=None, shrinkage=None, solver='svd', store_covariance=False, tol=0.0001)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.60 | 0.03 | 0.06 | 156 |
| food | 0.72 | 1.00 | 0.84 | 426 |
| host | 0.57 | 0.90 | 0.70 | 787 |
| endosymbiont | 0.99 | 0.93 | 0.96 | 359 |
| avg / total | 0.73 | 0.73 | 0.65 | 1728 |

Table A.1.9: QDA(priors=None, reg_param=0.0)

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.60 | 0.03 | 0.06 | 156 |
| food | 0.66 | 1.00 | 0.79 | 426 |
| host | 0.59 | 0.84 | 0.69 | 787 |
| endosymbiont | 0.99 | 0.93 | 0.96 | 359 |
| avg / total | 0.71 | 0.72 | 0.64 | 1728 |

Table A.1.10: GaussianNB()

| | precision | recall | f1-score | support |
|--------------|------------------|---------------|-----------------|----------------|
| unknown | 0.98 | 0.66 | 0.79 | 156 |
| food | 0.97 | 1.00 | 0.98 | 426 |
| host | 0.78 | 1.00 | 0.88 | 787 |
| endosymbiont | 0.98 | 0.99 | 0.99 | 359 |
| avg / total | 0.93 | 0.92 | 0.91 | 1728 |

Table A.1.11: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0)

A.2 eDicer

A.2.1 EUKARYOTE TRANSCRIPTOMES

eDicer Eukaryote Predicted Transcriptomes used:

- *Acanthamoeba castellanii* str. Neff
- *Aplanochytrium kerguelense* PBS07
- *Arabidopsis thaliana*
- *Aurantiochytrium limacinum*
- *Batrachochytrium dendrobatidis* JAM81
- *Bigelowiella natans*
- *Blastocystis hominis*
- *Bodo saltans*
- *Caenorhabditis elegans*
- *Capsaspora owczarzaki*
- *Chlamydomonas reinhardtii*
- *Chondrus crispus*
- *Ciona intestinalis*
- *Cryptococcus neoformans* var. grubii H99
- *Cryptosporidium parvum*
- *Cyanidioschyzon merolae*
- *Cyanophora paradoxa*
- *Dictyostelium discoideum*

- *Drosophila melanogaster*
- *Ectocarpus siliculosus*
- *Emiliania huxleyi* CCMP 1516
- *Entamoeba histolytica*
- *Fonticula alba*
- *Giardia intestinalis*
- *Guillardia theta*
- *Homo sapiens*
- *Hyaloperonospora arabidopsis*
- *Klebsormidium flaccidum*
- *Laccaria bicolor*
- *Monosiga brevicollis*
- *Mortierella verticillata* NRRL 6337
- *Mus musculus*
- *Naegleria gruberi*
- *Nannochloropsis gaditana*
- *Neurospora crassa* OR74A
- *Ostreococcus lucimarinus*
- *Perkinsus marinus*
- *Phaeodactylum tricornutum*
- *Physcomitrella patens*
- *Phytophthora ramorum*
- *Plasmodium falciparum*
- *Populus trichocarpa*
- *Reticulomyxa filosa*
- *Rozella allomycis* CSF55
- *Salpingoeca* sp. ATCC 50818

- *Schizosaccharomyces pombe*
- *Sphaeroforma arctica* jp610
- *Takifugu rubripes*
- *Tetrahymena thermophila macronucleus*
- *Thalassiosira pseudonana*
- *Thecamonas trahens* ATCC 50062
- *Toxoplasma gondii*
- *Trichomonas vaginalis* G₃
- *Trichoplax adhaerens*
- *Trypanosoma brucei*
- *Tuber melanosporum*
- *Ustilago maydis*
- *Vitrella brassicaformis* CCMP3155

A.2.2 BACTERIAL TRANSCRIPTOMES

eDicer Bacteria Predicted Transcriptomes used:

- *Acidimicrobium ferrooxidans* DSM 10331
- *Nocardia farcinica* IFM 10152
- *Frankia alni* ACN14a
- *Propionibacterium acidifaciens* DSM 21887
- *Kitasatospora setae* KM-6054
- *Bifidobacterium longum* NCC2705
- *Collinsella tanakaei* YIT 12063
- *Rubrobacter xylanophilus* DSM 9941
- *Conexibacter woeselii* DSM 14684
- *Moorella thermoacetica* ATCC 39073
- *Aquifex aeolicus* VF5

- *Persephonella marina* EX-H1
- *Desulfurobacterium thermolithotrophum* DSM 11699
- *Hydrogenobacter thermophilus* TK-6
- *Chthonomonas calidirosea* T49
- *Fimbriimonas ginsengisoli* Gsoil 348
- *Armatimonadetes bacterium* CSP1-3
- *Bacteroides fragilis* 3_1_12
- *Flavobacterium psychrophilum* JIPo2_86
- *Salinibacter ruber* DSM 13855
- *Chlorobium tepidum* TLS
- *Chlorobium luteolum* DSM 273
- *Chlorohherpeton thalassium* ATCC 35110
- *Prosthecochloris aestuarii* DSM 271
- *Ignavibacterium album* JCM 16511
- *Melioribacter roseus* P3M-2
- *Phycisphaera mikurensis* NBRC 102666
- *Isosphaera pallida* ATCC 43644
- *Blastopirellula marina* DSM 3645
- *Chlamydia trachomatis* 434/Bu
- *Waddlia chondrophila* WSU 86-1044
- *Lentisphaera araneosa* HTCC2155
- *Opitutus terrae* PB90-1
- *Akkermansia muciniphila* ATCC BAA-835
- *Anaerolinea thermophila* UNI-1
- *Caldilinea aerophila* DSM 14535
- *Chloroflexus aurantiacus* J-10-fl
- *Dehalococcoides mccartyi* 195

- *Ktedonobacter racemifer* DSM 44963
- *Thermomicrobium roseum* DSM 5159
- *Gloeobacter violaceus* PCC 7421
- *Nostoc punctiforme* PCC 73102
- *Acaryochloris* sp. CCME 5410
- *Synechococcus elongatus* PCC 6301
- *Synechococcus* sp. JA-3-3Ab
- *Chroococcidiopsis thermalis* PCC 7203
- *Pleurocapsa* sp. PCC 7327
- *Fischerella muscicola* PCC 7414
- *Calditerrivibrio nitroreducens* DSM 19672
- *Deferribacter desulfuricans* SSM1
- *Denitrovibrio acetiphilus* DSM 12809
- *Flexistipes sinusarabici* DSM 4947
- *Deinococcus radiodurans* R1
- *Truepera radiovictrix* DSM 17093
- *Meiothermus silvanus* DSM 9946
- *Thermus aquaticus* Y51MC23
- *Acidobacterium capsulatum* ATCC 51196
- *Fibrobacter succinogenes* subsp. *succinogenes* S85
- *Chitinivibrio alkaliphilus* ACh1
- *Bacillus subtilis* B7-s
- *Lactobacillus acidophilus* NCFM
- *Streptococcus suis* BM407
- *Halanaerobium praevalens* DSM 2228
- *Coprobacillus* sp. 3_3_56FAA
- *Mitsuokella multacida* DSM 20544

- *Anaerococcus hydrogenalis* ACS-025-V-Sch4
- *Thermoanaerobacter ethanolicus* JW 200
- *Cetobacterium* sp. ZOR0034
- *Fusobacterium nucleatum* 13_3C
- *Ilyobacter polytropus* DSM 2926
- *Leptotrichia wadei* Fo279
- *Sebaldella termitidis* ATCC 33386
- *Nitrospira defluvii*
- *Thermodesulfovibrio islandicus* DSM 12570
- *Leptospirillum* sp. Group I
- *Caulobacter crescentus* CB15
- *Agrobacterium tumefaciens* WRT31
- *Hyphomonas* sp. 25B14_1
- *Parvularcula bermudensis* HTCC2503
- *Wolbachia pipiensis* wAlbB
- *Sphingobium* sp. AP49
- *Acetobacter pasteurianus* 386B
- *Bordetella avium* 197N
- *Neisseria bacilliformis* ATCC BAA-1200
- *Azoarcus* sp. BH72
- *Nitrosomonas eutropha* C91
- *Methylobacillus flagellatus* KT
- *Aeromonas hydrophila* NJ-35
- *Escherichia coli* str. K-12
- *Pseudomonas aeruginosa* PAO1
- *Xanthomonas oryzae* ATCC 35933
- *Alcanivorax* sp. DG881

- *Vibrio fischeri* SR5
- *Shewanella baltica* BA175
- *Bdellovibrio bacteriovorus* HD100
- *Desulfarculus baarsii* DSM 2075
- *Myxococcus xanthus* DZ2
- *Desulfovibrio vulgaris* DP4
- *Geobacter sulfurreducens* KN400
- *Campylobacter jejuni* 10186
- *Helicobacter pylori* 35A
- *Nautilia profundicola* AmH
- *Nitratirductor salsuginis* DSM 16511
- *Sulfurovum lithotrophicum*
- *Brachyspira innocens* ATCC 29796
- *Leptospira interrogans* serovar Bim str. P2529
- *Borrelia burgdorferi* B31
- *Spirochaeta thermophila* DSM 6578
- *Treponema denticola* ATCC 33520
- *Synergistes* sp. 3_1_syn1
- *Aminomonas paucivorans* DSM 12260
- *Dethiosulfovibrio peptidovorans* DSM 11002
- *Anaerobaculum mobile* DSM 13181
- *Pyramidobacter piscolens* W5455
- *Acholeplasma laidlawii* PG-8A
- *Mycoplasma mycoides* subsp. *mycoides* SC str. PG1
- *Kosmotoga olearia* TBF 19.5.1
- *Petrotoga mobilis* SJ95
- *Fervidobacterium islandicum*

- *Caldisericum exile* AZM16co1
- *Desulfurispirillum indicum* S5
- *Dictyoglomus thermophilum* H-6-12
- *Elusimicrobium minutum* Pei191
- *Gemmatimonas aurantiaca* T-27
- *Gemmatimonadetes bacterium* KBS708
- *Nitrospina gracilis* 3/211
- *Mari profundus ferrooxydans* PV-1
- *Thermodesulfatator atlanticus* DSM 21156
- *Thermodesulfobacterium thermophilum* DSM 1276
- *Caldithrix abyssi* DSM 13497

A.2.3 ARCHAEA TRANSCRIPTOMES

eDicer Archaea Predicted Transcriptomes:

- *Acidianus hospitalis* W1
- *Acidilobus saccharovorans* 345-15
- *Aciduliprofundum boonei* T469
- *Aeropyrum pernix* K1
- *Archaeoglobus fulgidus* DSM 4304
- *Caldisphaera lagunensis* DSM 15908
- *Caldivirga maquilingensis* IC-167
- *Candidatus Korarchaeum cryptofilum* OPF8
- *Candidatus Nitrosopelagicus brevis*
- *Cenarchaeum symbiosum A*
- *Desulfurococcus fermentans* DSM 16532
- *Ferroglobus placidus* DSM 10642
- *Ferroplasma acidarmanus* fer1

- *Fervidicoccus fontis* Kam940
- *Halalkalicoccus jeotgali* B3
- *Haloarcula hispanica* ATCC 33960
- *Halobacterium salinarum* R1
- *Haloferax mediterranei* ATCC 33500
- *Halogeometricum borinquense* DSM 11551
- *Halomicrombium mukohataei* DSM 12286
- *Halopiger xanaduensis* SH-6
- *Haloquadratum walsbyi* DSM 16790
- *Halorhabdus utahensis* DSM 12940
- *Halorubrum lacusprofundi* ATCC 49239
- *Haloterrigena turkmenica* DSM 5511
- *Halovivax ruber* XH-70
- *Hyperthermus butylicus* DSM 5456
- *Ignicoccus hospitalis* KIN4 I
- *Ignisphaera aggregans* DSM 17230
- *Metallosphaera sedula* DSM 5348
- *Methanobacterium formicum* DSM 3637
- *Methanobrevibacter ruminantium* M1
- *Methanobrevibacter* sp. AbM4
- *Methanocaldococcus fervens* AG86
- *Methanocella conradii* HZ254
- *Methanococcoides burtonii* DSM 6242
- *Methanococcus maripaludis* S2
- *Methanocorpusculum labreanum* Z
- *Methanoculleus bourgensis* MS2
- *Methanofollis liminatans* DSM 4140

- *Methanohalobium evestigatum* Z-7303
- *Methanohalophilus mahii* DSM 5219
- *Methanolobus psychrophilus* R15
- *Methanomethylovorans hollandica* DSM 15978
- *Methanoplanus limicola* DSM 2279
- *Methanopyrus kandleri* AV19
- *Methanoregula formicicum* SMSP
- *Methanosaeta thermophila* PT
- *Methanosalsum zhilinae* DSM 4017
- *Methanosarcina acetivorans* C2A
- *Methanospaera stadtmanae* DSM 3091
- *Methanospaerula palustris* E1-9c
- *Methanospirillum hungatei* JF-1
- *Methanothermobacter marburgensis* str. Marburg
- *Methanothermococcus okinawensis* IH1
- *Methanothermus fervidus* DSM 2088
- *Methanotorris igneus* Kol 5
- *Nanoarchaeum equitans* Kin4-M
- *Natrialba magadii* ATCC 43099
- *Natrinema pellirubrum* DSM 15624
- *Natronobacterium gregoryi* SP2
- *Natronococcus occultus* SP4
- *Natronomonas pharaonis* DSM 2160
- *Nitrosopumilus maritimus* SCM1
- *Picrophilus torridus* DSM 9790
- *Pyrobaculum arsenaticum* DSM 13514
- *Pyrococcus yayanosii* CH1

- *Pyrolobus fumarii* 1A
- *Salinarchaeum* sp. Harcht-Bsk1
- *Staphylothermus hellenicus* DSM 12710
- *Sulfolobus islandicus* M.14.25
- *Thermococcus barophilus* MP
- *Thermofilum pendens* Hrk 5
- *Thermogladius cellulolyticus* 1633
- *Thermoplasma acidophilum* DSM 1728
- *Thermoplasmatales archaeon* BRNA1
- *Thermoproteus uzonensis* 768-20
- *Thermosphaera aggregans* DSM 11486
- *Vulcanisaeta distributa* DSM 14429