Computational analysis of eukaryotic metatranscriptomes from diverse marine environments

Andrew Toseland

Primary Supervisor: Professor Vincent Moulton

Secondary Supervisor: Dr. Thomas Mock

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia
School of Computing Sciences
Norwich, United Kingdom

August, 2013

Abstract

Phytoplankton are photosynthetic microbes that form the basis of the marine food web and are estimated to produce over half of all oxygen in the Earth's atmosphere. Recent advances in high-throughput DNA sequencing technologies have allowed scientists to sample the set of genes actively transcribed from communities of microbes *in-situ*. This set of transcripts (the metatranscriptome) provides a snapshot of actively transcribed genes at the time of sampling, and can provide insights into microbial metabolism and their relationship with their environment. In this thesis we present the computational analysis of eukaryotic phytoplankton metatranscriptome data sampled from representative marine environments; the simulation of metatranscriptome data for benchmarking computational tools; and analysis carried out on a newly sequenced eukaryotic phytoplankton genome.

Transcripts affiliated with ribosomal proteins and associated with translation dominated in all but the Equatorial Pacific metatranscriptome sample. Hierarchical clustering of the metatranscriptome samples by taxa produced two groups: the diatom dominated and the alveolate dominated. However, clustering by Gene Ontology terms clustered the samples by environment type (tropical, temperate and polar), producing a gradient of translation-associated transcripts which increased as the *in-situ* temperature of the samples decreased. A strong correlation ($R = 0.9$) was detected between the relative proportion of transcripts associated with temperature and the *in-situ* temperature. Laboratory experiments on model diatom species under control conditions confirmed that as the *in-situ* temperature decreases, these model diatoms produce more transcripts and consequently more ribosomal proteins.

A translational efficiency experiment demonstrated that the rate of translation decreased under low temperatures for a model diatom species. This suggested that the increased production of ribosomes acts as a compensatory mechanism under low temperatures. As more ribosomes require more phosphate-rich rRNAs we hypothesised that this could have an impact on biogeochemical cycles (E.g. the Redfield ratio of Nitrate (N) to Phosphate (P)). This was modelled by our collaborators from the University of Exeter, who produced a global phytoplankton cell model of resource allocation. They showed how the N:P ratio differs across latitudinal temperature zones and predicted the impact of increasing temperature on global N:P.

Acknowledgements

Firstly I would like to thank my supervisors Vincent Moulton and Thomas Mock for their guidance and support. I would also like to thank the Earth & Life Systems Alliance (ELSA) for funding my PhD. In addition, I would like to thank Simon Moxon for encouraging me to apply for a PhD in computational biology, and for his time and patience during the course of my studies. Finally, I would like to thank my parents Albert and Elizabeth, and my brother Matthew for their love and support over the last four years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The focus of the research presented in this thesis is the development of computational pipelines for the analysis of metatranscriptome data from eukaryotic phytoplankton. This has ranged from the analysis of environmental transcriptome data, sampled *in-situ* from phytoplankton communities, to simulating metatranscriptome data and benchmarking computational tools, and finally the analysis of transcript sequences from specific phytoplankton species. The research described in this thesis has involved collaborations with other groups within the UEA and other institutes. A brief overview of the thesis is provided below detailing my main contributions to each project.

In **Chapter 2** we provide background information about metatranscriptomics. We discuss the importance of this relatively new approach to studying microbial metabolism and some of its applications. We then go on to describe the group of microbial organisms that provide the focus of our analysis, the eukaryotic phytoplankton and their ecological significance. We then describe the relevant biological processes involved in metatranscriptomics and the current generation of high-throughput sequencing technologies employed to generate metatranscriptome sequence data. Finally, we describe the different stages of computational analysis involved in metatranscriptome analysis pipelines and provide an overview of some of the widely used computational tools and resources.

In **Chapter 3** we describe the computational pipeline that we implemented for the analysis of metatranscriptome sequence data from marine eukaryotic phytoplankton communities. We also detail follow up laboratory experiments carried out to confirm our findings and discuss the biological significance of the results. Sampling and sample preparation was performed by groups headed by Dr. Thomas Mock (UEA), Dr. Gareth Pearson (University of Algarve), and Dr. Klaus Valentin (Alfred Wegener Institute). The follow up laboratory experiments were carried out by members of Dr. Thomas Mock's group. Finally, the phytoplankton cell modelling analysis was performed by Dr. Stuart Daines and colleagues from the University of Exeter. My contribution was the implementation of the computational pipeline for the processing and analysis of all metatranscriptome sequence data.

In **Chapter 4** we describe an assessment of sequence processing methods on simulated metatranscriptome data. Due to a lack of simulated metatranscriptome data sets for benchmarking purposes, we also implemented a novel method to simulate data sets for different levels of taxonomic diversity. We also assessed the accuracy of protein domain annotation on sequences after the application of different processing methods. All work in this chapter was carried out by myself. Dr. Simon Moxon of the Genome Analysis Center, Norwich, helped to conceptualise and design the analysis.

In **Chapter 5** we describe the analysis of allelic variants in the eukaryotic phytoplankton *Fragilariopsis cylindrus*. By performing a comparative analysis with a sexually reproducing phytoplankton species, we tested the hypothesis that the cessation of sexual reproduction in *Fragilariopsis cylindrus* has increased allelic divergence, allowing it to adapt to an extreme polar environment. The transcriptome of *Pseudonitzschia multistriata* was sequenced by the Joint Genome Institute (JGI) for Dr. Mariella Ferrante of the Stazione Zoologica Anton Dohrn in Napoli. Quality filtering of *Pseudonitzschia multistriata* transcriptome sequences was performed by Dr. Remo Sanges, also of Napoli. Interpretation and analysis of the nucleotide divergence results was carried out by Dr. Mark McMullan and Dr. Cock Van Oosterhout of the School of Environmental Sciences at UEA. My contributions were the assembly of *Pseudonitzschia multistriata* sequences, the processing of *Fragilariopsis cylindrus* alleles, the detection of homologous alleles and the implementation of scripts to calculate nucleotide divergence.

Finally, in **Chapter 6** we discuss the work detailed in this thesis and provide some thoughts regarding extensions and directions this research could take in the future.

Chapter 2

Background

2.1 Summary

In this chapter we introduce metatranscriptomics and discuss the importance and applications of this approach. We provide a brief biological background into an important group of microbes, the phytoplankton, which are the focus of analysis for chapters 3 and 5; the processes of translation and transcription and current high-throughput sequencing platforms. We then discuss the various stages of a typical metatranscriptome project and summarise the current state of the art of available computational tools for processing, determining taxonomic affiliations, predicting function and providing comparative and statistical analyses of high-throughput sequence data.

2.2 Metatranscriptomics

"Although invisible to the naked eye, prokaryotes are an essential component of the Earth's biota. They catalyze unique and indispensable transformations in

the biogeochemical cycles of the biosphere, produce important components of the earths atmosphere, and represent a large portion of lifes genetic diversity". [Whitman et al., 1998]

It is estimated that approximately 99% of prokaryotes cannot be cultured under laboratory conditions [Amann et al., 1995], which greatly limits our understanding of microbial diversity, population composition and gene expression to a tiny minority of model organisms. However, recent advances in high-throughput DNA sequencing technologies have enabled scientists to sample, sequence and analyse community microbial genomic DNA (metagenomics) or mRNA (metatranscriptomics).

Metagenomics allows for the analysis of community composition and the discovery of novel organisms and genes [Fernández-Arrojo et al., 2010]. The largest metagenomic study to date, the Global Ocean Sampling (GOS) expedition, circumnavigated the globe, sampling microbial communities. The subsequent analysis produced over 6 million novel protein sequences [Rusch et al., 2007], [Yooseph et al., 2007] vastly expanding our knowledge of marine microbes. However, it was restricted to detailing the set of genes that could potentially be expressed and it tells us nothing about about actual gene expression levels [Moran, 2010]. In order to study community gene expression at the time of sampling a metatranscriptomic approach is required. By sequencing and analysing actively transcribed genes from a population of microbes we can get a real-time snapshot of community gene expression, see figure 2.1 for an overview.

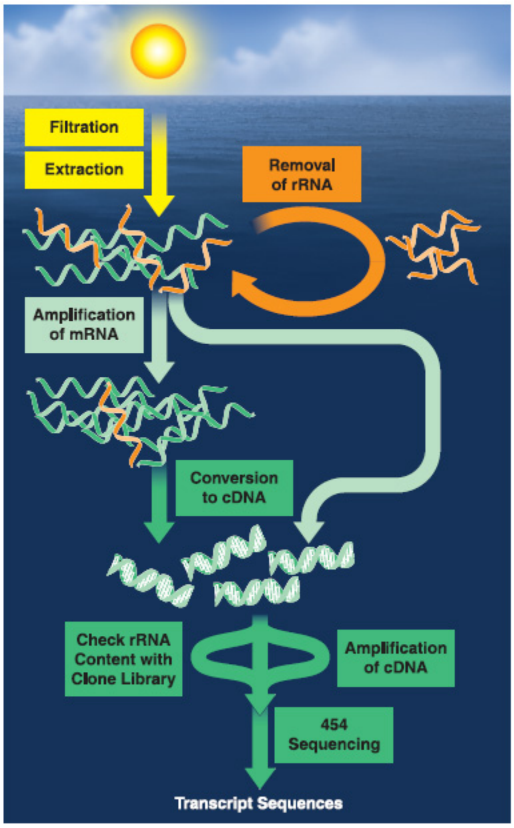


Figure 2.1: Typical protocol for sampling and sequncing of a marine metatranscriptome. Samples are size filtered, messenger RNA (mRNA) is then isolated and amplified before being converted into complimentary DNA (cDNA) for sequencing. Figure reproduced from [Moran, 2010].

Since the largest proportion of prokaryotes are found in soil and marine habitats [Whitman et al., 1998], it is unsurprising that the majority of metatranscriptome projects to date have focused on soil [Leininger et al., 2006], [Bailly et al., 2007], [Urich et al., 2008], [Baldrian et al., 2011], [Damon et al., 2012] and marine [Gilbert et al., 2008], [Frias-Lopez et al., 2008], [Gifford et al., 2010], [Lesniewski et al., 2012] microbial communities. Another key area of interest is the analysis of microbial communities present in the intestines of higher organisms [Poroyko et al., 2010], [Booijink et al., 2010], [Xiong et al., 2012]. Historically, the majority of such projects have focused on prokaryote communities. Recently however, a small number of projects have emerged to address eukaryotic microbial communities such as the eukaryotic proportion of phytoplankton [John et al., 2009], [Marchetti et al., 2012].

While the tools and methods employed to analyse metatranscriptome data vary greatly and depend to some extent on the type of data and aims of the project, metatranscriptome analyses can be broken down into a series of discrete stages: sampling, sequencing, sequence processing, determination of taxonomic composition, functional annotation, comparative and statistical analyses. These are addressed in more depth in later sections.

Most early metatranscriptome projects employed pyrosequencing [Gilbert et al., 2008], [Urich et al., 2008], [Gifford et al., 2010], a sequencing by synthesis method for determining the order of nucleotides in a DNA fragment (see section 2.5). However, short read platforms such as SOLiD and Illumina can now produce many times more sequence data at a lower per base cost - albeit at a cost of shorter reads which require assembly. Many recent projects have employed these methods, either solely Illumina [Qi et al., 2011], [Mason et al., 2012] or taken a hybrid approach using an assembled 454 backbone and mapping SOLiD reads to the backbone to provide a more accurate quantitative analysis [Marchetti et al., 2012].

Generally, functional annotation of microbial metatranscriptomes identify the majority of transcripts as being involved in fundamental processes such as biosynthesis and energy generation [Moran, 2010]. While this is hardly revelatory, the real power of metatranscriptomics lies in its ability to detect novel biocatalysts from unculturable organisms [Warnecke and Hess, 2009] and to perform comparative analyses of gene expression profiles - whether temporal [Poretsky et al., 2009], spatial [Stewart et al., 2012] or in regard to changing environmental conditions [Marchetti et al., 2012], [Mason et al., 2012] or during important biological processes such as phytoplankton blooms [Gilbert et al., 2008].

Studies providing a combined metagenome and metatranscriptome have shown that only a relatively small proportion of the metagenome is actively expressed [Frias-Lopez et al., 2008], and that the most abundant species are not necessarily the most active in terms of gene expression [Gilbert et al., 2008], [Stewart et al., 2012]. Also, as mRNA has a short half-life [Bernstein et al., 2002] compared to protein and not all transcripts are guaranteed to be translated into protein [Moran, 2010], the level of transcription of a gene reflected in a metatranscriptome does not necessarily reflect the abundance of protein and care must be taken when interpreting metatranscriptome expression profiles.

One shortcoming of projects to date is that a relatively small proportion of metatranscriptome sequences return matches to protein databases. Even using the comprehensive NCBI REFSEQ database, only between 13-37% of all sequences returned matches [Frias-Lopez et al., 2008], [Poretsky et al., 2009], [Gifford et al., 2010], [Qi et al., 2011]. Although previous studies contained a high degree of non-coding RNA, protocols for the removal of rRNA from prokaryotic samples have been developed [Stewart et al., 2010] and for eukaryotic samples mRNA can be specifically targeted. Therefore, the lack of homologous sequences may be partly due to the length and quality of sequences currently available [Wommack et al., 2008]. However, as previously stated only a small proportion of microbes have been successfully cultured and sequenced and it seems likely that a lack of reference genomes is the root cause.

2.3 Phytoplankton

The term phytoplankton is a compound word derived from the Greek - *phytos* (plant) and *planktos* (wandering or drifting), literally meaning wandering plant. It is an umbrella term for a diverse group of unicellular, photosynthetic organisms including both eukaryotes and prokaryotes, inhabiting the pelagic or upper layer of marine and freshwater environments worldwide.

Across coastal and open ocean systems and even polar sea ice, phytoplankton form the basis of the marine food chain. Powered by sunlight, they take up inorganic compounds directly from their environment such as water and carbon dioxide and synthesise organic carbon, releasing oxygen as a by-product. The process of producing organic compounds carbon dioxide is known as primary production and phytoplankton are estimated to be responsible for approximately 50% of global primary production [Field et al., 1998]. Other nutrients are required for phytoplankton growth (primarily nitrate and phosphate) and, when sufficient light and nutrients are available, phytoplankton populations grow exponentially forming blooms visible from space. The phytoplankton are grazed upon by microscopic animals (zooplankton), which in turn are consumed by fish, which are eaten by larger predatory fish which are finally eaten by humans. Due to their short life-cycle, rapid growth and the speed with which they react to changes in their environment (nutrient levels, temperature etc.), phytoplankton are widely studied as indicators of environmental conditions [McCormick and Cairns, 1994] and there has been much interest in how predicted global warming could affect phytoplankton populations [Boyce et al., 2010]. Other important areas of phytoplankton research are their potential use as biofuels [Bozarth et al., 2009] and attempting to reduce anthropogenic carbon dioxide levels by artificial iron fertilization of phytoplankton [Denman, 2008].

The most abundant phytoplankton are the prokaryotic cyanobacteria, but in terms of standing stock (biomass per unit volume of water) the most important groups are the larger eukaryotes: the Bacillariophyceae (diatoms), Dinophyceae (dinoflagellates), Haptophyceae (coccolithophorids and prymnesiomonads) and Cryptophyceae (cryptomonads) [Parsons et al., 1983]. The most diverse group of eukaryotic phytoplankton are the diatoms, consisting of some 200,000 species [Armbrust, 2009]. These can be put into two major groups: the (bipolar or multi-polar) centric (circular or cylindrical) and the (raphid or araphid) pennate (rod-shaped) diatoms (see figure 2.2). All diatoms are distinguished by their intricate silicate cell wall or frustule and overlapping valve structure. The way in which these intricate silicate structures are formed is of great interest to researchers in the field of nanotechnology [Bradbury, 2004], [Gordon et al., 2009].

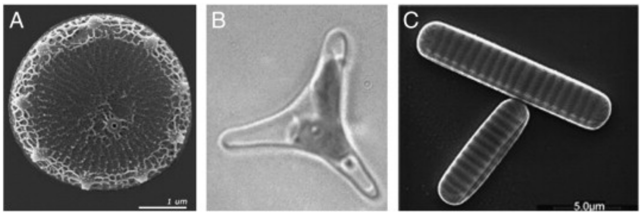


Figure 2.2: A) The centric diatom *Post-classical pseudonana*. B) The pennate diatom *Phaeodactylum tricornutum*. C) The polar pennate diatom *Fragilariopsis cylindrus*. Image reproduced with permission from [Smith et al., 2012].

Diatoms are estimated to have emerged between 190 [Sims et al., 2006] and 250 [Sorhannus, 2007] million years ago and have a complex evolutionary history. Around 1.5 billion years ago a eukaryotic organism assimilated the genetic material (known as endosymbiosis) from a cyanobacteria, which formed the chloroplasts of plants, red and green algae. A second endosymbiosis occurred some 500 million years later when a red algae was engulfed by another eukaryote to create the diatom [Armbrust, 2009]. Diatoms are secondary endosymbionts containing a chimeric mixture of animal, plant and bacterial genetic material.

Despite so much interest in diatoms, only two species have had their complete genomes sequenced to date; these are the centric diatom *Thalassiosira pseudonana* [Armbrust et al., 2004], and the pennate diatom *Phaeodactylum tricornutum* [Bowler et al., 2008]. At the time of writing two other diatom genomes - the polar diatom *Fragilariopsis cylindrus* [Institute, a] and *Pseudonitzschia multiseries* [Institute, b] are in the draft stage.

2.4 Genetic material

The Central Dogma of Molecular Biology as first stated by Francis Crick in 1958 [Crick, 1958] and clarified in 1970 [Crick et al., 1970] deals with the flow of genetic information. Essentially, it states that the normal transfers of genetic information are: 1) DNA to DNA - DNA self-replicating during cell division for example; 2) DNA to RNA - where the DNA of a gene is *transcribed* into complementary RNA; 3) RNA to protein - after transcription RNA is *translated* into protein. There are other transfers of genetic information, but this is the fundamental flow of genetic information for most organisms. So, for a gene to be expressed, DNA is transcribed into RNA which is then translated into protein.

Proteins are essential components of cellular activity for all organisms and they perform a variety of functions: enzymes control chemical reaction rates; structural proteins such as keratin form hair and nail tissue; they can also be hormones and influence cell signalling [Klug and Michael, 1997]. Proteins are formed from linear sequences of amino acids (composed from a set of 20), which fold into a three dimensional shape. Each of the 20 different amino acids has a specific set of properties, the aggregate effect of which determine the properties of the protein to be formed [Alberts et al., 2002]. These properties effect the eventual shape of the protein and the type of molecules it can interact with thereby influencing the function of the protein.

The instructions for creating these protein structures are encoded into the DNA of an organism's genes. With eukaryotes, these genes are contained in the cell nucleus on chromosomes - discrete structures made of DNA tightly wrapped around histone proteins, and the complete set of chromosomes is called the genome. Bacterial genomes are generally a single, circular chromosome free floating in the cell.

DNA is made of two complimentary strands of nucleotides. Nucleotides consist of a sugar-phosphate backbone and one of the four nucleobases: adenine, cytosine, guanine or thymine, abbreviated to **A**, **C**, **G** and **T** [Klug and Michael, 1997], see figure 2.3. Different bases have an affinity with each other, allowing them to form a hydrogen bond. For example, adenine binds to thymine (and vice-versa), and cytosine to guanine (and vice-versa) [Klug and Michael, 1997]. This allows nucleotides on opposite strands of DNA to bind (each matching pair of bases is called a base pair) and join the two strands together. This also means that one strand is the reverse complement of the other.

When we talk about genome size, we talk about the total number of base pairs (bp) in terms of kilobases (1 thousand bases), megabases (1 million bases) and gigabases (1 billion bases). For example the first organisms to have their complete genome sequence determined were viral and consisted

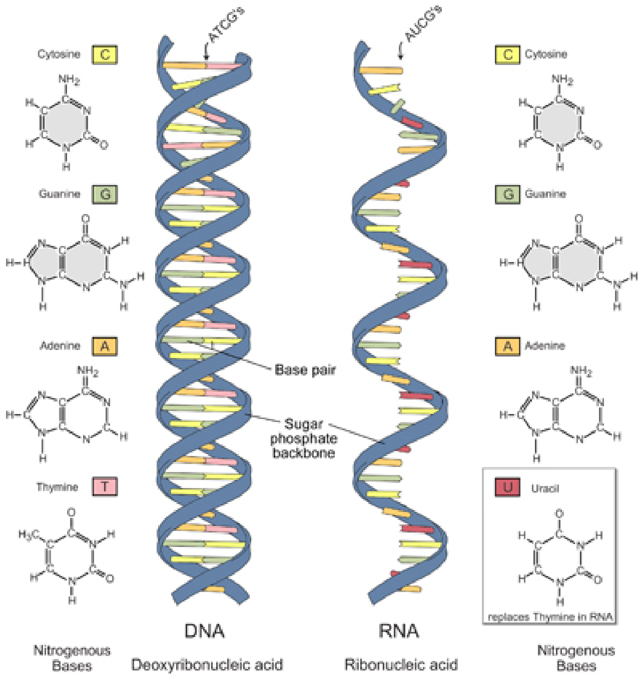


Image adapted from: National Human Genome Research Institute.

Figure 2.3: Left: structure of double stranded deoxyribonucleic acid (DNA). Right: structure of single stranded ribonucleic acid (RNA). Figure reproduced from http://www.tutorvista.com/biology/types-of-dna-and-rna.

of 3-5 Kb [Fiers et al., 1976], [Fiers et al., 1978]. The genome of the first sequenced bacteria - *Haemophilus influenzae* contained a total of 1.8 Mb [Fleischmann et al., 1995] and the human genome contains approximately 3.2 Gb [Venter et al., 2001].

As mentioned above, for a gene to be expressed, it must first be transcribed into RNA and then translated into protein. The double-stranded DNA is first separated into two strands, the coding strand and it's complement, the template strand. In transcription however, only the template strand is copied into complementary messenger RNA (mRNA, also known as a transcript). As the mRNA is complementary (is anti-parallel) to the template strand it is therefore identical (except for **U**racil replacing **T**hymine) to the corresponding section of the coding strand, see figure 2.3.

Not all of the DNA in a gene codes for protein; for example, eukaryotic genes are composed of coding areas (exons) and non-coding areas (introns) [Klug and Michael, 1997]. At this stage, both are transcribed, as well as untranslated regions (UTRs) at both ends of the gene which don't code for protein but serve other important roles, such as providing binding sites for ribosomes. Some post-transcriptional modification also occurs; more specifically, the mRNA has its introns spliced out, and the exons are then ligated together. A cap is added to the 5' end and a stretch of adenines called a poly-A tail is added to the 3' end [Klug and Michael, 1997]. The mRNA is now a mature mRNA and can leave the cell nucleus and enter the cytoplasm to be translated. This is summarised in figure 2.4 below.

In the cytoplasm, a protein producing factory called a ribosome attaches to the mRNA at a binding site in the 5' UTR and translates the strand of mRNA into a chain of amino acids which form a protein [Klug and Michael, 1997]. The mRNA is fed through the ribosome 3 nucleotides at a time, each tri-nucleotide sequence encodes for an amino acid and is called a codon. One of the 64 possible codons, AUG which encodes methionine, is a start codon, and signifies the start of the translation process. Other RNA molecules called transfer RNAs (tRNAs) enter the ribosome carrying the required amino acids, which pair to mRNA codons through a complementary tri-nucleotide adaptor called an anti-codon. If the tRNA anti-codon complements the current mRNA codon, the two bond and the amino acid carried by the tRNA is released and added to a growing chain of amino acids called a polypeptide. Translation terminates when one of three stop codons is reached, the polypeptide is released, folds and

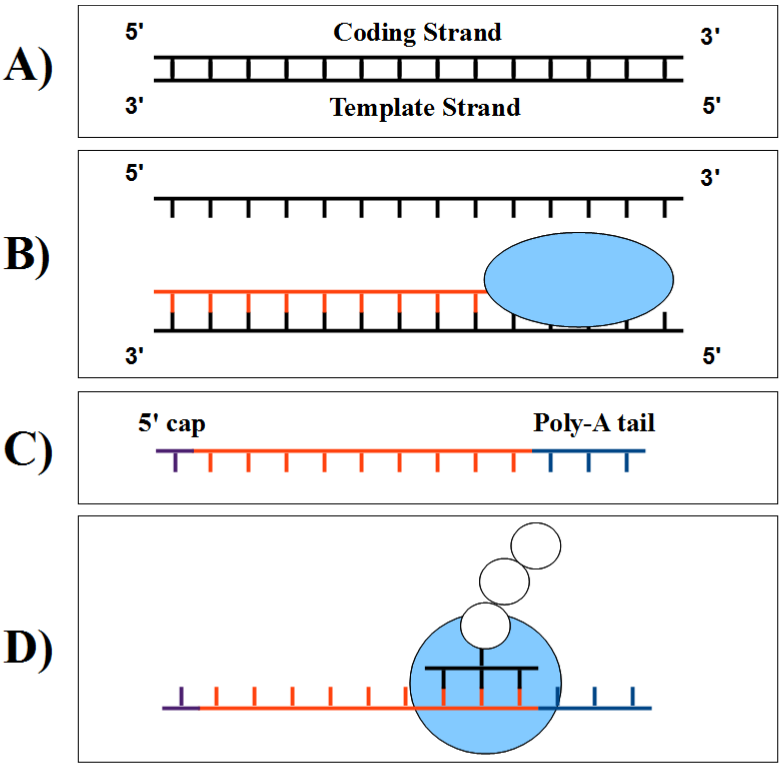


Figure 2.4: A) Double stranded DNA. B) Transcription - the double stranded DNA is divided into two strands: the coding strand and the template strand. The enzyme polymerase (blue oval) incorporates nucleotides into a complementary copy of the template strand, this is a single stranded messenger RNA (mRNA). C) Mature mRNA with 5' cap and 3' poly-A tail. D) Translation - The mRNA is translated in a ribosome (blue circle) into chains of amino acids which combine to form proteins.

becomes a protein or a sub-unit of a protein.

2.5 Next generation sequencing technologies

In order to study an organisms' genome: to analyse its structure; investigate known and novel genes; to perform comparative genomics; to look at it's evolutionary history through phylogenomics; or to study the set of genes expressed (the transcriptome) under varying conditions, it is necessary to extract and determine the precise sequence of nucleotides contained within the DNA or RNA molecule(s). This process is called DNA sequencing.

The first generation of DNA sequencing techniques were developed in the 1970s: Frederick Sangers' chain-termination method [Sanger et al., 1977] and Maxam-Gilbert chemical sequencing [Maxam and Gilbert, 1977]. Due to the relative ease and reliability of Sanger sequencing and the use of radioactive materials in the Maxam-Gilbert method, Sanger sequencing became the *de-facto* method of automated DNA sequencing. With Sanger sequencing, a DNA molecule is fragmented into smaller pieces and cloned in colonies of *Escherischia coli*. These fragments are extracted and amplified by multiple rounds of Polymerase chain reaction (PCR), where the DNA is heat separated into two separate strands (denatured) and each strand incorporates free floating deoxynucleotides (dNTPs) to create a double stranded duplicate of the original fragment. The final round of PCR includes the incorporation of flourescently labelled dideoxynucleotides (ddNTPs – using a different colour for each of the four bases), which terminate the extension of a DNA strand. Finally, in a process called capillary electrophoresis, the negatively charged DNA fragments are pulled towards a positive charge and a laser light to stimulate the fluorescent labelled terminator nucleotides. The fragments' speed is relative to their length and so short fragments move towards the positive charge faster, when they pass the laser light the fluorescent label is activated and the light emitted is measured and the nucleotide determined (See [Church, 2006], [Shendure and Ji, 2008] for a full description).

Sanger sequencing is still in use and produces relatively long ($\sim$900bp), high accuracy sequences. It is however, a slow and expensive process with low throughput. By comparison, the new generation of sequencing technologies, can be run in a massively parallel fashion to produce orders of magnitude more data at a fraction of the cost and time. For example, the current Sanger sequencing platform can produce around 100 kilobases of data per run at a cost of ∼$2,400 per million bases sequenced, whereas second generation platforms such as 454 pyrosequencing [Margulies et al., 2005], Illumina and SOLiD can produce approximately 0.7, 600 and 120 gigabases of data at a cost of $10, $0.07 and $0.13 per million bases sequenced respectively (see [Liu et al., 2012] for a comparison of sequencing platforms). However, it should be noted that sequencing errors are more likely with these technologies. For example, pyrosequencing has problems determining homopolymeric regions and Illumina sequences tend to be biased towards GC rich regions [Dohm et al., 2008] and computational techniques are required to process the data.

We will focus here on briefly describing two of the most widely used second generation platforms - 454 pyrosequencing and Illumina methodologies. For both methods, the DNA to be sequenced is sheared into smaller fragments, amplified through either emulsion PCR (pyrosequencing) or bridge PCR (Illumina). With pyrosequencing, a nucleotide wash is added to the amplified sequences a single base at a time and using a combination of enzymes the release of pyrophosphate is detected and the base determined [Ronaghi, 2001], see figure 2.5 for an overview of 454 pyrosequencing. With Illumina sequencing flourescently labelled deoxynucleotides are added in cycles, one of which is incorporated and the identity of the incorporated base is determined by exciting the fluorescent tag with a laser.

The final product of a sequencing run is a set of sequences files; generally in FASTA format and a separate file of quality scores for 454, and FASTQ combined sequence and quality scores for Illumina, see figure 2.6. FASTQ files

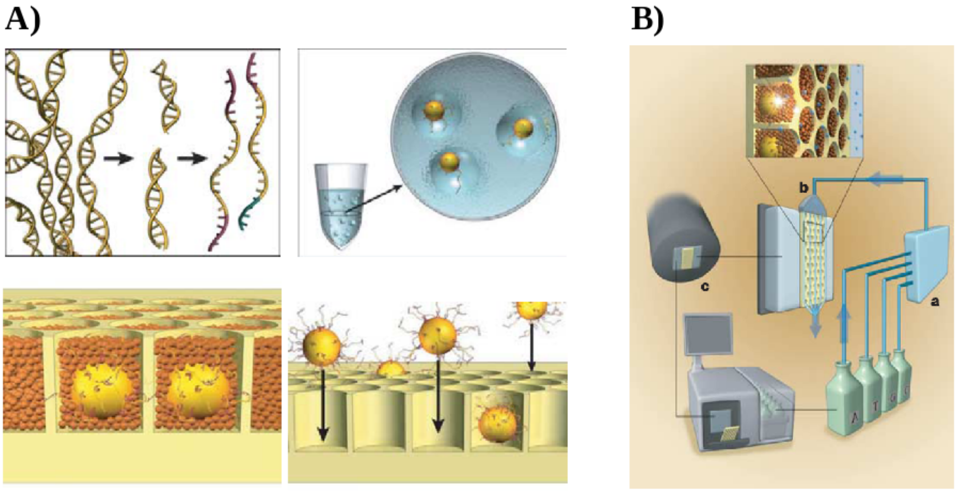
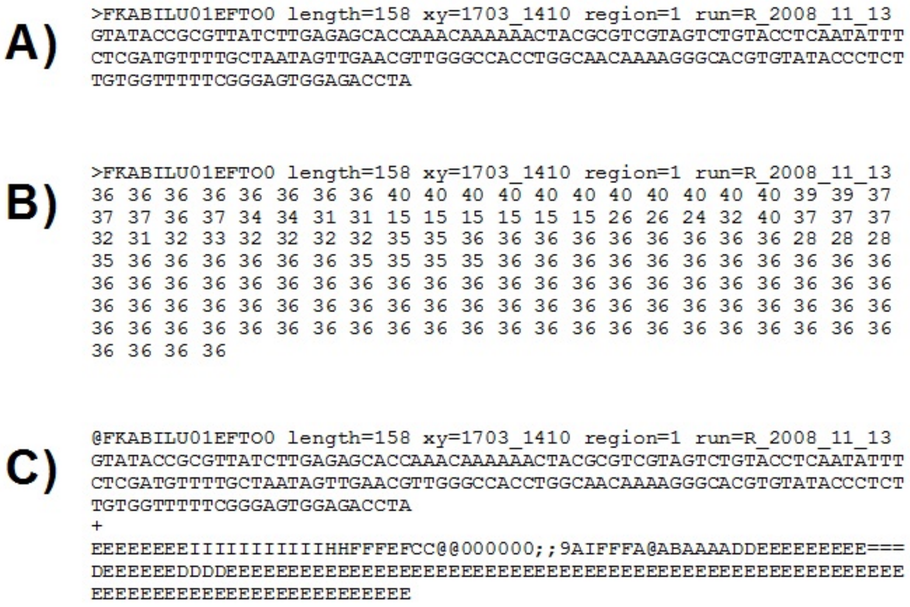


Figure 2.5: Overview of 454 pyrosequencing. A) Clockwise from top left. DNA shearing: longer DNA molecules (e.g. genomic DNA) are sheared into fragments of a few hundred base pairs. Emulsion PCR: adaptor sequences are ligated to the fragments allowing them to bind to capture beads which are placed into an oil and water mixture containing amplification reagents. Bead loading: the amplified beads are loaded into a sheet of microscopic wells called a PicoTiterPlate. Well packing: each wel is then filled with tiny beads, tightly packing in the amplified beads. B) a) Individual nucleotides are sequentially flowed over the well. b) As nucleotides are incorporated into a template strand on an amplifiaction bead pyrophosphate is released. c) The release of light is detected by a camera and the base determined. Figure reproduced from [Margulies et al., 2005].

consist of 4 lines per sequence, the first line of each sequence begins with an @ and contains a unique identifier for each sequence, optionally followed by details of the sequencing run. The second line contains the sequence of nucleotides itself, the third line contains an optional description and the fourth line contains the ASCII encoded sequence quality scores (phred scores), one score per base, in the same order as the sequence. The quality scores represent the probability that a particular base is incorrect and calculated by the following formula: $Q = -10 * log_{10}(p)$. So, for example, a quality score of 15 represents a 0.0316% probability that the base is erroneous.



A)

```
>FKABILU01EFTO0 length=158 xy=1703_1410 region=1 run=R_2008_11_13
GTATACCGCGTTATCTTGAGAGCACCAAACAAAAAACTACGCGTCGTAGTCTGTACCTCAATATTT
CTCGATGTTTTGCTAATAGTTGAACGTTGGGCCACCTGGCAACAAAAGGGCACGTGTATACCCTCT
TGTGGTTTTTCGGGAGTGGAGACCTA
```

B)

```
>FKABILU01EFTO0 length=158 xy=1703_1410 region=1 run=R_2008_11_13
36 36 36 36 36 36 36 36 40 40 40 40 40 40 40 40 40 40 40 39 39 37
37 37 36 37 34 34 31 31 15 15 15 15 15 15 26 26 24 32 40 37 37 37
32 31 32 33 32 32 32 32 35 35 36 36 36 36 36 36 36 36 36 28 28 28
35 36 36 36 36 36 36 35 35 35 35 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36
```

C)

```
@FKABILU01EFTO0 length=158 xy=1703_1410 region=1 run=R_2008_11_13
GTATACCGCGTTATCTTGAGAGCACCAAACAAAAAACTACGCGTCGTAGTCTGTACCTCAATATTT
CTCGATGTTTTGCTAATAGTTGAACGTTGGGCCACCTGGCAACAAAAGGGCACGTGTATACCCTCT
TGTGGTTTTTCGGGAGTGGAGACCTA
+
EEEEEEEEIIIIIIIIIIIHHFFFEFCC@@000000;;9AIFFFA@ABAAAADDEEEEEEEEE===
DEEEEEEDDDDEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEE
```

Figure 2.6: A) Nucleotide sequence in FASTA format. B) Separate quality file for FASTA sequence. C) DNA sequence from (A) converted into FASTQ format.

FASTA files are plain text files containing a set of sequences. A ">" denotes the beginning of each sequence followed by a unique identifier, the subsequent lines contain the actual sequence of nucleotides itself from 5' to 3'. The quality file contains a set of phred equivalent scores ranging from 0 to 40 for each base of each sequence, representing the probability that the base is an overcall [Brockman et al., 2008], that is, an extra nucleotide inserted into the sequence.

As previously mentioned, second generation sequencing platforms represent a huge increase in the throughput of sequence data while reducing both the cost and time involved in their production. The downside is that these technologies produce shorter sequences than Sanger; up to 700bp for the latest 454 FLX Titanium platform and over100bp paired reads with Illumina HiSeq [Liu et al., 2012].

2.6 Sequence processing

Before any downstream analysis is performed, the raw sequence data must first be quality filtered. As 'omics samples consist of hundreds of thousands or even millions of sequences and the taxonomic and functional annotation process is computationally intensive, the aim is to remove as much unwanted data as possible. Artifacts from the sequencing process also must be removed and poor quality reads filtered out. In addition, for metatranscriptomics, decisions need to be made about the omission or inclusion of rRNA sequences and whether or not to assemble the sequences or use clustering as a data reduction strategy, that is, to remove identical or near identical sequences from subsequent analyses.

During metatranscriptome library preparation, primer sequences are annealed to mRNAs to facilitate reverse-transcription into double-stranded cDNA. Also, during sequencing, primers and adaptors are attached to the cDNA strands for PCR amplification. These regions must be detected and trimmed away as they could affect downstream stages. With 454 sequencing in particular, data sets may contain artificial duplicates [Gomez-Alvarez et al., 2009] - identical, or near-identical copies of genuine sequence data which, if not addressed could bias downstream quantification of sequence annotation. Approaches based on sequence clustering have been developed to detect and remove artificial duplicates [Niu et al., 2010]. However, metatranscriptome data sets contain a high degree of redundancy. This is because we only sequence a subset of genomes - the transcribed regions and the more highly expressed a transcript, the more likely it is that there are genuine duplicates in the data. It is therefore commonly assumed that the majority of duplicates in metatranscriptome data are natural and reflect real transcript abundance [Niu et al., 2010].

Though the methods and parameters vary, other common filtering stages include trimming by quality score; removing sequences under a certain length (for example the MG-RAST metagenomics processing pipeline [Meyer et al., 2008] only uses reads >75bp); removal of low complexity sequences (for example reads containing an unusually high degree of a single base [Hewson et al., 2009]) or containing ambiguous bases (Ns). In addition to mRNAs, metatranscriptome samples may contain other types of RNA that are not translated into proteins. Examples of these non-coding RNAs are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) which are estimated to constitute between 95 and 97% of total RNA in bacteria [Rosenow et al., 2001], between 80 and 85% in yeast [Von Der Haar, 2008] and ∼80% in mammalian cells [Lodish et al., 2000]. Ribosomal RNA makes up the greatest proportion of total RNA and some purification steps are required to specifically target mRNA. These rRNA sequences can be detected by alignment against curated databases of rRNA sequences such as Silva [Pruesse et al., 2007], RDP [Maidak et al., 2001], GreenGenes [DeSantis et al., 2006] or through HMM methods like RNAmmer [Lagesen et al., 2007]. If sufficient numbers of rRNA sequences are available they can be extracted from the sequence pool and used for taxonomic classification (see section 2.8), otherwise, they should be omitted from downstream analyses.

Several sequence processing pipelines have been developed offering a range of filtering options. The MG-RAST server [Meyer et al., 2008] provides options to filter reads by quality score, length and ambiguous bases. SeqTrim [Falgueras et al., 2010] offers adaptor removal, quality trimming, and contamination detection. Prinseq [Schmieder and Edwards, 2011], offers both an online and a stand-alone processing pipeline providing a comprehensive set of filtering options for high-throughput data including: primer and adaptor removal, quality trimming, length filtering, rRNA detection, duplicate detection and also graphical and statistical reports of the data and the filtering results.

Once filtered, in order to reconstruct individual genomes or transcriptomes from the mixed pool of an 'omics dataset, the data must be assembled. The aim of assembly is to detect overlapping regions between sequences and join them into longer contiguous stretches of nucleotides (contigs). There are two main methods of assembly: De-Bruijn graphs and the overlap layout consensus (OLC) method. Both have their origins in graph theory, see [Li et al., 2012] for a full description. De-Bruijn graph based assemblers such as Velvet [Zerbino and Birney, 2008], SOAP denovo [Li et al., 2010], and the transcriptome assembler Trinity [Grabherr et al., 2011] are generally used for short-read Illumina assemblies. For longer 454 or Sanger reads, OLC assemblers are preferred such as MIRA [Chevreux et al., 2004] or 454's proprietary assembler Newbler [Margulies et al., 2005]. Assembling a single genome or species specific transcriptome can be problematic due to missing or repetitive regions, variable coverage, sequencing biases or contaminations. When the data to assemble originates from multiple organisms assembly is even more difficult; 'omics data sets often assemble poorly into short fragmented contigs and the possibly exists of sequences from two or more similar organisms to merge into chimeric contigs. Therefore many authors choose to omit assembly [Kunin et al., 2008]. However assemblers have been developed specifically designed to overcome the problems inherent in 'omics data such as MetaVelvet [Namiki et al., 2011] and metaIDBA [Peng et al., 2011].

If however, the aim of an analysis is to provide a quantitative overview of a microbial population rather than genome/gene reconstruction, sequence clustering is a useful alternative to assembly, and can be used to remove redundancy thereby reducing the size of the data set before performing homology searches [Thomas et al., 2012]. Clustering tools like CD-HIT [Li and Godzik, 2006], Uclust [Edgar, 2010] or as part of the MG-RAST pipeline [Meyer et al., 2008] use fast, heuristic methods to remove redundancy from sequence data by grouping sequences sharing a defined degree of similarity into clusters. The longest or representative sequence from each cluster can be used for downstream, computationally intensive analyses rather than the entire group of sequences.

2.7 Sequence homology

As stated in section 2.4, the sequence of amino acids that constitute a protein determine its properties and thereby influence it's function. Thus, a common approach in predicting the function of an unknown sequence is by inference through homology. An unknown or novel sequence can be aligned to sequences with a known, empirically determined function and, if it is deemed to be homologous, then we can predict that the novel sequence has a similar function. This allows us to search for conserved protein domains within DNA sequences by comparing them to a database of sequences and predict function by detecting highly similar or homologous sequences.

The majority of tools for producing sequence alignments are based on dynamic programming [Eddy, 2004]; techniques such as the Needleman-Wunsch and Smith-Waterman algorithms, although computationally expensive, detect the optimal global (alignment across the entirety of the sequences) and local (similar regions shared by sequences) sequence alignments respectively. Two widely used methods for detecting sequence homology are the Basic

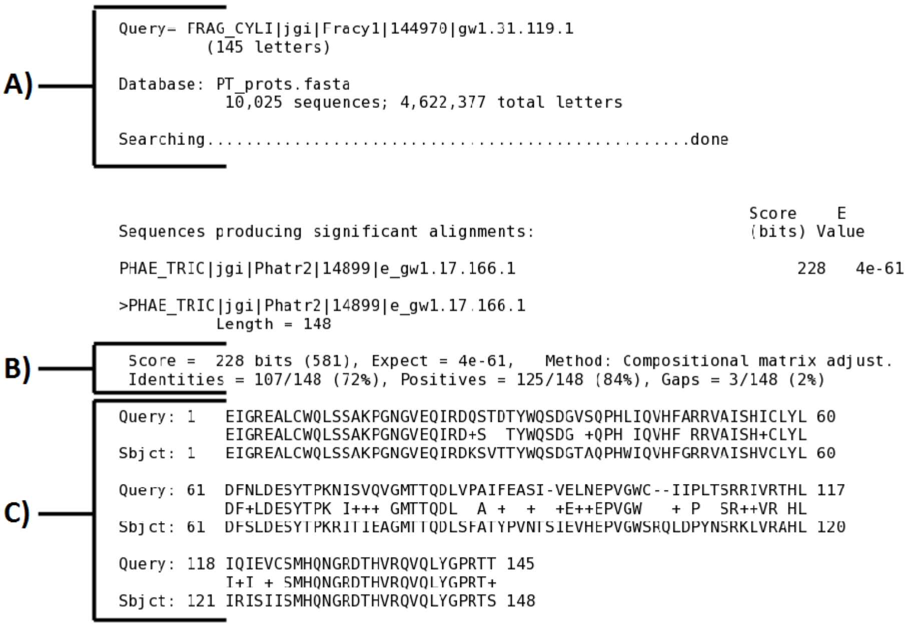
Local Alignment Search Tool (BLAST) [Altschul et al., 1997] and HMMER [Eddy et al., 2009].

2.7.1 BLAST

The most widely used tool for performing sequence homology searches is BLAST [Altschul et al., 1997], the Basic Local Alignment Search Tool. Rather than aligning two sequences over their entire length, (global alignment) BLAST looks for conserved regions shared by the two sequences (local alignment). It takes an heuristic approach to sequence alignments. To align one sequence (the query) to another, (the target) the query sequence is first split into short sub-sequences called words. If a word can be aligned to the target it forms a seed alignment, the seed alignment is extended as far as possible in either direction to find the optimal matching sequence region(s) known as High-scoring Segment Pairs (HSPs).

BLAST comes in several variations allowing for many types of alignment. The most commonly performed are alignments of nucleotide to nucleotide (BLASTN), amino acid to amino acid (BLASTP) and nucleotide to amino acid sequences (BLASTX), which compares the 6 frame translation of the nucleotide sequence to the amino acid sequence. When a seed match is found and optimally extended it is assessed by using a substitution matrix to produce a score for the alignment; BLAST uses the BLOSUM matrix by default [Henikoff and Henikoff, 1992]. Each pair of amino acids between the query and target sequence is assessed. If they are identical the score is increased by the appropriate value from the matrix. If they do not match but share similar properties - for example, hydrophilic amino acids are more closely related to each other than to hydrophobic residues - the score is increased by a smaller amount.

If the bases do not share similar properties, or there is a gap then a penalty is incurred. The final score is normalised over the sequence length to produce a standardised quality score for the alignment called the bit-score. Another value is produced for each alignment called the e-value or expect value. The e-value is a measure of probability representing the likelihood of an alignment of the bit-score returned occurring by chance in the query, see figure 2.7 for an example BLAST report.



```
Query= FRAG_CYLI|jgi|Fracy1|144970|gw1.31.119.1
         (145 letters)

Database: PT_prots.fasta
           10,025 sequences; 4,622,377 total letters

Searching..................................................done


                                                              Score    E
Sequences producing significant alignments:                   (bits) Value

PHAE_TRIC|jgi|Phatr2|14899|e_gw1.17.166.1                        228   4e-61

>PHAE_TRIC|jgi|Phatr2|14899|e_gw1.17.166.1
          Length = 148

 Score =  228 bits (581), Expect = 4e-61,   Method: Compositional matrix adjust.
 Identities = 107/148 (72%), Positives = 125/148 (84%), Gaps = 3/148 (2%)

Query: 1   EIGREALCWQLSSAKPGNGVEQIRDQSTDTYWQSDGVSQPHLIQVHFARRVAISHICLYL 60
           EIGREALCWQLSSAKPGNGVEQIRD+S  TYWQSDG +QPH IQVHF RRVAISH+CLYL
Sbjct: 1   EIGREALCWQLSSAKPGNGVEQIRDKSVTTYWQSDGTAQPHWIQVHFGRRVAISHVCLYL 60

Query: 61  DFNLDESYTPKNISVQVGMTTQDLVPAIFEASI-VELNEPVGWC--IIPLTSRRIVRTHL 117
           DF+LDESYTPK I+++ GMTTQDL  A +  +  +E++EPVGW   + P  SR++VR HL
Sbjct: 61  DFSLDESYTPKRITIEAGMTTQDLSFATYPVNTSIEVHEPVGWSRQLDPYNSRKLVRAHL 120

Query: 118 IQIEVCSMHQNGRDTHVRQVQLYGPRTT 145
           I+I + SMHQNGRDTHVRQVQLYGPRT+
Sbjct: 121 IRISIISMHQNGRDTHVRQVQLYGPRTS 148
```

Figure 2.7: Example BLAST report. A) Summary section showing the name of a query sequence, the databases to be searched and the alignment progress. B) Alignment statistics including e-value and bit-score. C) The sequence alignment itself. The alignment contains 3 lines: the query sequence on top, the matching database sequence at the bottom and the alignment sequence or high-scoring segment pair (HSP) in the middle. The coordinates of the matching sequences are shown at the begininning and end of the alignment.

While BLAST is very much a standard bioinformatics tool, it is limited in several ways: firstly BLAST only returns the best matches found in the reference database used and, due to the relatively low number of sequenced organisms available, most reference databases are biased towards a small group of model organisms. For functional analysis this may be sufficient to accurately detect close homologs, but for determining the taxonomic origins of a sequence the most significant match returned may be taxonomically distant from the real source organism [Koski and Golding, 2001].

2.7.2 HMMER

HMMER [Eddy et al., 2009] uses Hidden Markov Models (HMMs) to apply position-specific models to sequence alignments. An HMM is a trained probabilistic model consisting of a sequence of state transitions, known as a Markov Chain. The model consists of a series of states and a set of associated transition probabilities reflecting the probability of progressing from the current state to a different (or the same) state (see figure 2.8). Each state also has a set of emission probabilities. These represent the likelihood of the model emitting a value from a discrete language set at this state. The set of emissions is the observable output of an HMM, but the underlying rules contained in the state transitions for producing the output are said to be hidden.

HMMs are used by protein databases including Pfam [Finn et al., 2010] and TigrFam [Haft et al., 2003]; for rRNA detection and classification tools like RNAmmer [Lagesen et al., 2007]; for gene or reading frame detection programs like GeneWise [Birney et al., 2004] and ESTscan [Iseli et al., 1999] (where gene features such as start and stop codons, introns, exons are used to identify genic regions); and for determining the taxonomic affiliations of sequences e.g. Carma [Krause et al., 2008].

With Pfam for example, an HMM is produced for each protein family by

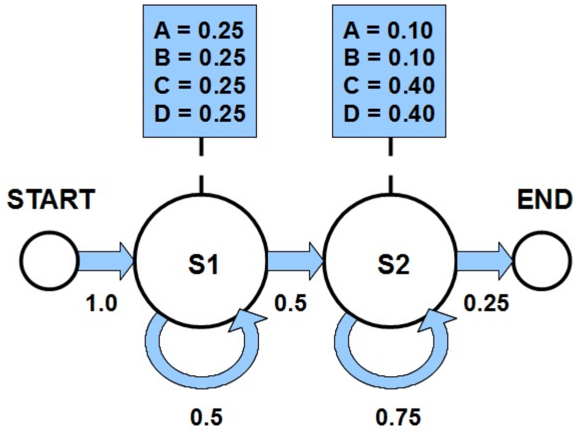


Figure 2.8: Example HMM - Showing states (circles), state transitions (solid arrows) and their associated transition probabilities, and emission probabilities (rectangles). This Markov chain proceeds from the start state through states S1 and S2 to an end state. The probability of moving from one state to another is represented by the transition probability, and the likelihood of producing each of the four possible outputs (A-D) is reflected in the set of emission probabilities.

creating a multiple sequence alignment of homologous protein sequences. The consensus structure and variance of the protein sequences - highly conserved residues, frequency of insertions and deletions, can then be modelled and the residue emission and state transition probabilities calculated. For a Pfam HMM, the possible states are: match - the two residues are identical (or at least share similar properties); insertion - the residue is present in the query sequence but not the consensus; deletion - the residue is present in the consensus but absent from the query. The residue emissions or observations are the set of 20 amino acids.

Essentially what is being modelled here is the probability of each residue (amino acid) in a sequence belonging to the model based on the previous residue. Query sequences are compared to the Pfam HMMs using HMMER [Eddy et al., 2009] and an accumulative probability value is produced representing the likelihood of the query sequence having been produced by the HMM and therefore having a similar function to other proteins of the protein family being modelled. Certain nucleotides or amino acids will be more highly conserved than others in proteins of a similar function. The strength of HMMs is that, unlike a BLAST search they can weight certain residues according to their level of conservation. However, querying HMMs is computationally intensive, and they require extensive sets of training data to calibrate the models.

2.8 Taxonomic classification

One of the main aims of 'omics is to identify and quantify the taxonomic composition of a set of sequence data. A wide variety of applications have been designed for this purpose and new tools and methods are being published every year. This section provides an overview of some of the most widely used tools at the time of writing.

Broadly speaking, the approaches for determining the taxonomic affiliations can be categorised as either similarity-based or composition-based. For many years, the gold standard for determining taxonomic diversity was by the targeted sequencing and classification through similarity searches against curated databases of 16s rRNA, a highly conserved gene ubiquitous in prokaryotic organisms (or 18s rRNA in eukaryotes). Many computational tools exist for the classification of 16s sequences such as UniFrac [Lozupone and Knight, 2005], Mothur [Schloss et al., 2009] and Arb [Ludwig et al., 2004]. However, this approach may lead to biased results due to variations in 16s rRNA copy numbers or from PCR bias [Liu et al., 2011] and many alternative approaches have been developed.

One such tool is MLTreeMap [Stark et al., 2010], which aligns sequences against a small set of highly conserved marker genes (16s/18s, RuBisCO etc.). Those that match a marker gene are incorporated into a multiple sequence alignment containing the marker gene sequences and placed on a phylogenetic tree using maximum likelihood. While MLTreeMap can taxonomically place sequences containing marker genes relatively accurately (>85% accuracy at the phylum level), only a tiny proportion of environmental sequences will contain these marker genes and MLTreeMap could only classify ∼1% of sequences from a metagenomic sample.

Metaphyler [Liu et al., 2011] uses a set of 31 marker genes, based on the Amphora pipeline [Wu et al., 2008], but can be extended to include all available genomes. Classification of input sequences is determined by a BLAST similarity approach using automatically learned settings from the reference database. Metaphyler is fast, can provide an accurate overview of the taxonomic composition of environmental sequence data (>90% precision at all taxonomic levels) and can potentially identify novel taxa. However, as with other marker gene approaches, only the subset of sequences containing the marker genes can be classified. Carma3 [Gerlach and Stoye, 2011] can use either BLAST or HMM results (e.g. from Pfam) and works on the assumption that different protein families will have different mutation rates but that the rate of mutation among members of the same family is consistent. Sequences are taxonomically placed based on the bit-scores of reciprocal BLAST searches between the region of sequence used for the initial alignments.

The metagenome analyser MEGAN is a Java application with a range of features for taxonomic, functional and comparative analyses of one or more data sets [Huson et al., 2007], [Mitra et al., 2010], [Mitra et al., 2011].

It uses a BLAST pre-processing stage, where sequences are compared to a reference database such as the NCBI non-redundant protein database [Pruitt et al., 2007], UniProt [Bairoch et al., 2005] or Silva [Pruesse et al., 2007]. The BLAST results are loaded into MEGAN, along with a set of user defined parameters, (bit-score cut-off, number of allowable hits per sequence etc.) and the sequences are mapped onto the NCBI taxonomy tree (built from all organisms contained in GenBank) using a Lowest Common Ancestor (LCA) Algorithm. For example if a sequence has matches to two taxa a and b, then it is placed on the taxonomic tree at the lowest node that contains both a and b as descendants. MEGAN is very user friendly and performs well compared to best BLAST approaches such as that employed by the MG-RAST [Meyer et al., 2008] pipeline.

Composition-based tools work on the principle that the frequency of oligonucleotides (short region of single-stranded DNA generally less than 20 nucleotides) in an organisms' genome contain a species-specific signature that can be used to characterise DNA fragments [Kariin and Burge, 1995], [Teeling et al., 2004a]. As the name suggests, Tetra [Teeling et al., 2004b] uses tetranucleotide usage to classify DNA fragments. First, the frequency of all 256 (4^4) tetranucleotides in a DNA fragment is calculated. These are then compared to the tetranucleotide frequencies of a set of reference genomes and the correlation coefficients calculated for each comparison. Phylopythia [McHardy et al., 2006] uses a multi-class support vector machine (SVM) trained on variable length oligonucleotides of reference genomes. While it performs with $>$80% and $>$90% accuracy for sequences of unknown and known origin respectively Phylopythia [McHardy et al., 2006], performance tails off drastically for fragments under 1 Kb in length. PhymmBL [Brady and Salzberg, 2009], is a hybrid classifier which combines nucleotide composition (Phymm) and similarity based methods (BLAST). The Phymm part of the tool builds interpolated Markov models (IMMS) from variable length oligomers (weighted according to frequency of occurrence) of a set of reference genomes. When used in conjunction with BLAST alignments, PhymmBL can classify reads as short as short as 100 bp with accuracy comparable to or greater than stand-alone BLAST or other composition-based methods such as PhyloPythia.

It is clear that there is as yet no 'one size fits all' taxonomic classifier. As with many bioinformatics tools there are trade-offs to be made between speed, accuracy and coverage. Similarity-based methods tend to work well with most length sequences, although accuracy and specificity will increase with sequence length [Wommack et al., 2008]. Most composition-based methods need longer sequences to determine a clear signal of oligonucleotide usage and perform poorly on sequences shorter than 1 Kb. Similarity-based methods can be biased by both under and over-represented taxa in the reference databases used [Piganeau, 2012], and often the closest BLAST match may not necessarily represent the closest phylogenetic match [Koski and Golding, 2001]. Marker gene approaches tend to have better performance but only allow classification of a tiny subset of sequences. Composition-based methods are reliant on having sufficient training data available to build reliable reference models [McHardy and Rigoutsos, 2007].

2.9 Functional annotation

A key stage in an 'omics analysis is determining the functional potential of a set of genes detected within a metagenomic sample or the functional activity in the transcripts of expressed genes in a metatranscriptome sample. The putative function of a gene or transcript sequence is inferred by comparisons of their amino-acid translations to one or more reference databases using either BLAST or HMM searches, sequences that produce matches within defined similarity or probability thresholds are deemed to be homologous. The functional annotation of the matching sequence is assigned as a predicted function.

The National Centre For Biotechnology Information (NCBI) provides two comprehensive reference databases, GenBank [Benson, 2011] and the non-redundant protein database RefSeq [Pruitt et al., 2012]. GenBank is built from genomic data, transcripts and environmental sequences from Genbank itself in the U.S., EMBL in the U.K. and the DNA Data Bank of Japan. At the time of writing GenBank contains over 150 Gb of nucleotide sequence from over 300,000 organisms. RefSeq is a more compact reference set of genomic, transcript and EST sequences. It is nonredundant at the species level, meaning that identical or highly similar sequences are represented by a single sequence. The current release contains over 4 billion amino acid sequences.

While these are among the most comprehensive reference databases available, the sequences are from a variety of sources and the quality of annotation is inconsistent, sometimes erroneous [Schnoes et al., 2009] and many sequences have no functional annotation at all, or are classifed as either predicted or hypothetical proteins based on similarity to existing sequences. The UniProt Knowledge Base (UniProtKB) [Bairoch et al., 2005] consists of two parts: Swiss-Prot and TrEMBL. Swiss-Prot contains highly accurate, [Schnoes et al., 2009] manually curated protein sequences with detailed information about the function and structure of the sequence, as well as information from relevant literature. TrEMBL consists of automatically annotated, and therefore less reliable sequences.

Another approach to predicting the function of a sequence is to attempt to cover the majority of protein coding sequences using representative sets of orthologous proteins from different organisms separated into families of proteins or functional groups. The key concept is that the majority of proteins are highly conserved among different species, as protein shape determines function and amino acid sequence determines protein shape, so major mutations in a sequence would likely disrupt the function of a gene. So, in determining the function of a sequence we do not necessarily need to compare it to all proteins from all organisms; the conserved consensus is suffcient.

The Pfam protein families database [Finn et al., 2010] is maintained by the Sanger Institute and uses manually curated and calibrated profile HMMs to represent families of homologous proteins. Sequences are queried against these probabilistic models to judge the likelihood that the sequence could have been produced by the model. The current Pfam database is built from protein sequences from UniProt [Bairoch et al., 2005], and covers around 80% of UniProt proteins in a set of around 13,500 protein families. It therefore provides high coverage of protein sequences with a small reference set. The manually curated models provide high quality, consistent annotation.

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [Ogata et al., 1999] is an hierarchical knowledge-base containing detailed, empirically validated information on metabolic pathways and the protein components and their interactions. Groups of orthologous proteins are represented by single KEGG orthologys (KO), sequences can be aligned with BLAST to KOs through the KAAS web server [Moriya et al., 2007] or against a local downloaded version of the database and partial or complete pathways reconstructed.

KEGG results are somewhat ambiguous as proteins may be involved in multiple interactions in multiple pathways. However, tools exist to reduce a set of KO results, such as MinPath [Ye and Doak, 2009], which uses a parsimony approach to provide the minimal set of pathways that explain a results set. To visualise the complete set of detected metabolic pathways, tools such as iPath [Letunic et al., 2008] allow for customisable pathway maps of interacting metabolic pathways.

The Gene Ontology (GO) [Ashburner et al., 2000] was conceived as a consistent, universal set of gene and gene product descriptions. Contained within a directed acyclic graph structure, GO terms are divided into three categories: Cellular Component, Molecular Function and Biological Process and provide detailed descriptions of genes and gene products in these contexts. It is rarely the case that we find a one-to-one mapping between genes and GO terms. Often a single sequence can be associated with multiple GO terms from different categories at different levels of detail. Thus, many tools have been developed to not only detect significantly different GO terms between data sets, but also to summarise and visualise reduced sets of GO terms. ReviGO [Supek et al., 2011] for instance reduces and plots lists of GO terms into clusters of terms based on a measure of semantic similarity; WeGO [Ye et al., 2006] enforces a tree-like hierarchy onto list of GO terms, allowing the user to define a level of detail for summary. Finally, it is possible to map annotations from other sources to GO terms; for example Pfam (http://www.geneontology.org/external2go/pfam2go), or NCBI sequences through Blast2GO [Conesa et al., 2005].

The NCBI Clusters of Orthologous Groups (COG) and euKaryotic Orthologous Groups (KOG) databases [Tatusov et al., 2003] contain groups of orthologous proteins representing prokaryotes and eukaryotes respectively. Each COG or KOG represents a highly conserved protein structure common to multiple species. For each database, a small representative set of organisms were chosen and conserved proteins detected by finding 3 way reciprocal BLAST hits between their genomes. Orthologs detected among at least 3 of the representative organisms are then subjected for manual curation and annotation using other data repositories such as GenBank. The eukaryote specific database provides a high level functional annotation, grouping sequences into 25 broad categories, however, it is built from just 7 species and has not been updated since it's inception.

2.10 Comparative analyses

As stated in section 2.2, the power of metatranscriptomics lies in it's ability to capture a real-time snapshot of community mRNA abundance allowing for comparative studies. After detecting taxonomic affiliations and predicting the function of sequences, the final and arguably most important stage of any metatranscriptomics analysis pipeline is to summarise, visualise and detect statistically significant differences – and similarities - between samples.

Generally what is reported is the number of sequences matching to certain taxa, or the percentage of reads or percentage of assignable reads. Common methods for detecting significantly different taxa or transcripts work using either Fisher's exact test or Chi-squared test [McDonald, 2009]. These methods use 2x2 contingency tables containing the proportion of sequences matching to taxa/gene X and the proportion of sequences not matching to X in two samples. A measure of probability (p-value) is calculated representing the likelihood of observing the table given the null hypothesis (i.e. that there is no difference in the abundance of X in the two samples). It should be noted that the Chi squared p-value is an approximation, but fast to calculate and suitable for large values, whereas Fisher's exact is more computationally intensive but more sensitive to small values (rare taxa/genes). During multiple tests, false positives can be introduced, dependent on the p-value cutoff used, hence p-value correction is required [White et al., 2009]. A common approach is to use Bonferonni correction, lowering the p-value according to the number of tests performed.

As described in section 2.9, various tools exist specifically for the visualisation of GO terms, many of which can be used for detection of enriched (significantly more abundant in one sample) GO terms. The majority of these use either Fisher's exact test or Chi-squared test (see [Sherman et al., 2009] for more details). The Metagenome Analyser MEGAN [Huson et al., 2007] introduced in section 2.8, originally a graphical user interface tool for the analysis and visualisation of taxonomic profiles through tree structures, has been extended to incorporate not only functional analysis but also comparative analyses of multiple samples. The approach they have implemented to detect differential expression or significantly different taxa is called the directed homogeneity test. It is based on Pearson's chi-square test (optionally Bonferonni corrected) and assesses not only individual nodes but all children of a node. Other contingency table approaches include XIPE [Rodriguez-Brito et al., 2006], which uses a difference of medians approach and STAMP [Parks and Beiko, 2010] which attempts to identify differences that are biologically significant and not just statistically significant.

Microbial organisms have a profound effect on their surrounding environment, but this relationship is not unilateral, environment conditions (temperature, light, nutrients etc.) effect both the behaviour - in terms of gene expression – and the population mix of microbial communitites. It is important to be able to contextualise the gene expression profile of a microbial community with regard to their surroundings and determine the impact of different environmental factors on community gene expression. The software package Primer-E [CLARKE, 1993] provides an array of statistical functions including multivariate statistical tools such as multidimensional scaling (MDS) and principal components analysis (PCA). For example, a canonical correlation analysis applied to metagenomic data from the GOS data set [Rusch et al., 2007], [Yooseph et al., 2007] showed that temperature, sunlight, oxygen and carbon dioxide played a larger role in determining the distribution of genes associated in various metabolic pathways than salinity and nutrients [Raes et al., 2011].

2.11 Discussion

In this chapter we have given an overview of metatranscriptomics, a powerful approach for analysing and comparing the community gene expression of unculturable microbial organisms from diverse environments. We have provided some background on the fundamental biology underpinning this approach; the importance of such projects and some of their applications. We have also given an overview of the different steps required for performing a metatranscriptomics analysis and discussed a selection of the most popular computational tools employed. There is no standard protocol for the computational analysis of metatranscriptome data; the approach employed is usually dependent on the project and its aims. In the next chapter we implement and apply a computational analysis pipeline to metatranscriptome data from eukaryotic phytoplankton communities from a range of environments.

Chapter 3

Computational analysis of eukaryotic phytoplankton metatranscriptomes

This chapter is adapted from ‘The impact of temperature on marine phytoplankton resource allocation and metabolism’, A. Toseland, S. Daines, J. Clark, A. Kirkham, J. Strauss, C. Uhlig, T.M. Lenton, K. Valentin, G.A. Pearson, V. Moulton, T. Mock, accepted for publication at Nature Climate Change, 29th July 2013.

3.1 Summary

This chapter describes the analysis of 454 sequence data from marine eukaryotic metatranscriptomes from distinct latitudinal temperature zones. This project was a collaboration between the Schools of Computing Sciences and Environmental Sciences at UEA, the Alfred Wegener Institute in Bremerhaven, the University of Algarve, and the University of Exeter. In this chapter we will focus on the computational analysis which was carried out in the School of Computing Sciences. For a breakdown of author contributions see Chapter 1.

3.2 Background

To better understand the effect that temperature has on the metabolism of eukaryotic phytoplankton we undertook a large-scale metatranscriptome, study sampling communities across a range of temperature zones. Phytoplankton are influenced by their ambient temperature in many ways, for example, their metabolism and diversity are dependent on their temperature optima for growth [Eppley, 1972] and on temperature driven physical constraints such as stratification and mixing for the supply of nutrients [Falkowski et al., 1998]. Although hotly debated, it has been suggested that global warming has already had a profound impact on phytoplankton standing-stock, reducing it by approximately 1% per year [Boyce et al., 2010]. Despite the significance of temperature for marine phytoplankton, especially in the context of anthropogenic global warming, we currently have a limited understanding of its impact on eukaryotic phytoplankton growth, metabolism, and community composition.

We sequenced the metatranscriptomes of eukaryotic phytoplankton communities across a latitudinal range of temperature zones. Samples were taken from all 3 major marine temperature zones: polar (Arctic (ARC) and Antarctic (ANT)), temperate (North Pacific (NPAC) and North Atlantic (NATL)) and tropical (Equatorial Pacific (EPAC)) (see figure 3.1). The aims of this analysis were to determine the taxonomic composition and transcriptional profile of each sample, to determine the differences and similarities between samples and finally to investigate the impact of environmental conditions on phytoplankton community transcript expression. The remainder of this chapter details the bioinformatic analysis of metatranscriptome data. We present a schematic overview of the bioinformatics analysis in figure 3.2. For completeness, we also include a brief description of follow-up molecular biology experiments performed in the Mock lab and discuss some of the results from a novel cellular resource allocation model developed at the College of Life and Environmental Sciences from the University of Exeter as part of this project.

3.3 Materials and methods

3.3.1 Sampling and sequencing

Phytoplankton cells were taken from water samples obtained during research vessel cruises, the samples were filtered to remove non-microbial matter and the samples were then frozen in liquid nitrogen. For the Antarctic sea-ice samples, ice cores were drilled and the samples melted in sea water before filtering and freezing. Next, total RNA was extracted from the samples, these were then purified to remove non eukaryotic mRNA. The mRNA samples were then reverse transcribed into double-stranded complimentary DNA (cDNA) for sequencing with the 454-GS-FLX and 454 Titanium platforms. For a full description of the materials methods used for sampling and sequencing see Appendices A.1 and

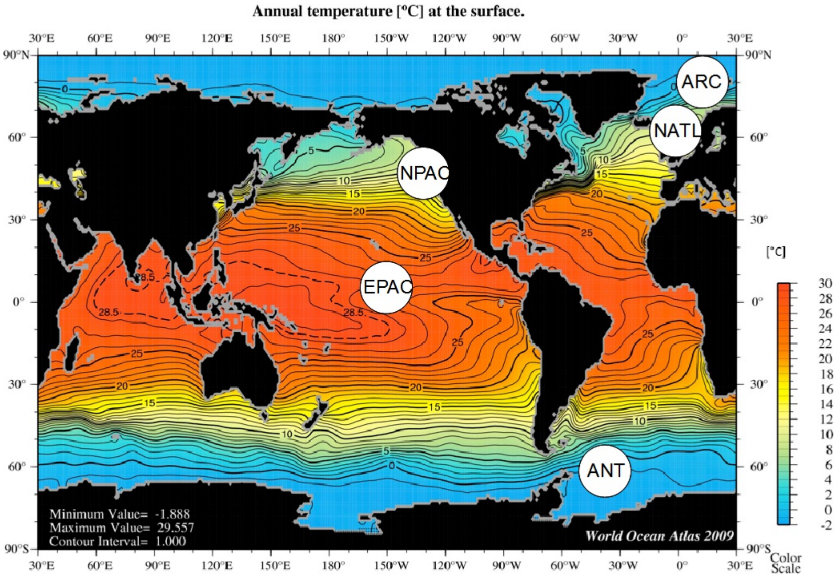


Figure 3.1: Sampling sites and surface ocean temperatures according to the World Ocean Atlas (2009). ARC: Arctic Ocean; NATL: North Atlantic Ocean; NPAC: North Pacific Ocean; EPAC: Equatorial Pacific; ANT: Southern Ocean.

A.2.

3.3.2 Sequence processing

As the quality of 454 sequences tends to deteriorate towards the 3' end, we performed clipping by quality score. Quality clipping was performed as in Marchetti et al. [Marchetti et al., 2012]. Using a single base sliding window, we trimmed each sequence from the 3' end until a base with a quality score of $\geq$14 was met. To identify potential sequencing artifacts, we clustered all sequences with CD-HIT-est [Li and Godzik, 2006] at 100% identity requiring 100% coverage of both sequences. We retained only the cluster representatives; cluster members (exact duplicates) were deemed potential artifacts and were omitted from further processing. To detect the 5' primer sequence we used the short sequence

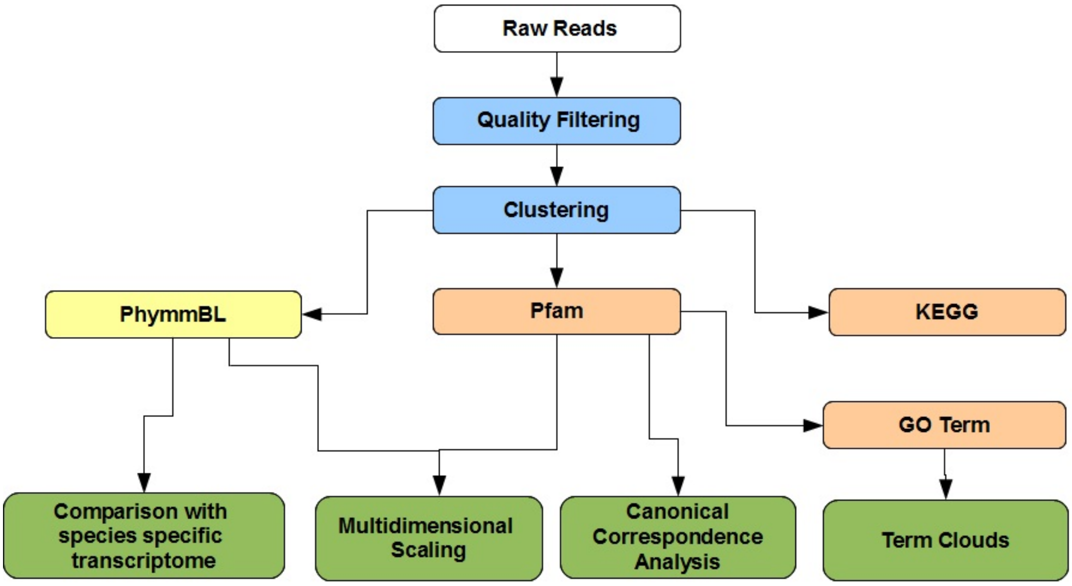


Figure 3.2: Workflow diagram of computational metatranscriptome analysis. Raw sequences (white) go through several processing stages (blue), before taxonomic (yellow), functional (pink) and finally comparative and statistical analyses (green).

alignment tool PatMan [Prüfer et al., 2008] allowing for up to 4 mismatches and 2 gaps. We used the match coordinates from PatMan to identify primer regions and these were removed using a custom BioPerl script. The 3', 17 base oligo-dt primer was identified using Dust [Kuzio et al., 2006] (word size 2, complexity value of 50) to get the coordinates of low complexity regions. We examined each identified region and, if it was of an appropriate length (≥15 bases) and composed of ≥75% adenine or ≥75% thymine, the region was trimmed from the sequence. We identified low complexity sequences using Dust [Kuzio et al., 2006]. Using default parameters, we ran all sequences through Dust, any low complexity regions were masked with Xs and the proportion of masked bases for each sequence was calculated. Any sequences comprising of ≥70% low complexity region were filtered out. Finally we removed any sequences that were less than 50 bp in length.

Despite specifically targeting eukaryotic mRNA by attaching oligo-dt primers to the poly-A region of transcripts, some non mRNA could have been present in the samples. John et al. [John et al., 2009] reported that ∼2% of sequences in a small scale eukaryotic metatranscriptome held significant similarity to ribosomal RNA (rRNA). In order to detect putative rRNA sequences we performed BLASTN [Altschul et al., 1997] searches (default settings, no complexity filtering) against both the large and small subunit databases of the Silva ribosomal database [Pruesse et al., 2007]. Sequences returning hits with bit scores ≥50 were deemed putative rRNA and we excluded them from further analysis.

The final processing stage was to cluster sequence sets to remove redundancy and speed up homology searches. We clustered each sample with CD-HIT-est at ≥95% overall identity and requiring ≥50% coverage of the representative sequence. We created a lookup table of cluster details (Cluster representative ID, Cluster size, Cluster member Ids) in order to scale the annotation results of cluster representatives accordingly.

To explore the composition of the samples and to detect broad-scale similarities and differences between samples, we pooled all 5 sequence sets together and clustered the amino-acid translations. We added an environment specific prefix to the unique identifier of sequences from all samples to denote their sample of origin and pooled them together. We then translated all sequences into their longest open reading frames (minimum length ≥10 amino acids) and clustered them with CD-HIT (≥90% overall identity, ≥50% coverage of the representative sequence). Using a custom Perl script, we examined the resulting clusters individually and appended the sequence ids of all cluster members to a list for each environment involved in that cluster. The resulting lists were used to create a sequence distribution Venn diagram in R using the venn function of the gplots package.

3.3.3 Taxonomic affiliations

To determine the taxonomic composition of the samples we used PhymmBL [Brady and Salzberg, 2009], a hybrid classifier which combines BLAST alignments with nucleotide composition based interpolated markov models. By default, PhymmBL uses bacterial and archaeal genomes from NCBI GenBank [Benson, 2011] as a reference. It is however, extensible and has been successfully applied to eukaryotic data [Brady and Salzberg, 2011]. We created a representative set of 44 eukaryotic organisms using genomes and ESTs (Expressed Sequence Tags) covering the major eukaryote groups but with a focus on algal species for this analysis (see Appendix A.3) for a list of organisms used and taxonomic labels. It was important to include out-groups (non algal species) to ensure that the analysis was not biased towards algal species and also to include any organisms that may have contaminated the samples during earlier stages. Genome sequences were downloaded from NCBI GenBank and JGI (with 4 exceptions: *Cyanidioschyzon merolae* from Cyanidioschyzon merolae Genome Project http://merolae.biol.s.u-tokyo.ac.jp/download; *Strongylocentrotus purpuratus* from Sea Urchin Genome Project http://www.hgsc.bcm.tmc.edu/projectspecies-o-Strongylocentrotus; *Danio rerio* from UCSC http://genome.ucsc.edu/cgi-bin/hgGateway?db=danRer5; and *Homo sapiens* from Genome Reference Consortium http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). EST sequences were downloaded from NCBI-dbEST and clustered with CD-HIT-est at 95% similarity to ensure non-redundancy of sequences.

We used taxonomic classifications for the PhymmBL configuration file from the NCBI taxonomy [Federhen, 2012] and AlgaeBase [Guiry and Guiry, 2011]. The sequence files and taxonomic details were added to PhymmBL in batch mode and IMMs created for each new organism. The cluster representative sequences of the metatranscriptome samples were run through the PhymmBL pipeline, and the results were filtered with a confidence score cutoff of $\geq$0.9 at the phylum level and scaled by the size of cluster.

3.3.4 Functional annotation

Pfam

We translated clustered representative sequences into all six reading frames (Min length 10 amino acids) and performed homology searches against the Pfam protein database [Finn et al., 2010] using pfam_scan.pl (Pfam-A only, default gathering thresholds used). The results were formatted with a custom Perl script and scaled according to cluster size.

Rarefaction curves

Rarefraction curves were produced with the online Rarefaction tool (http://www.biology.ualberta.ca/jbrzusto/rarefact.php#Calculator) using the Chao estimator of species richness. A list of raw totals for each detected Pfam domain were entered (plus the number of sequences providing no hits) and sampled at 50,000 sequence intervals.

KEGG

We identified KEGG pathways [Ogata et al., 1999] for cluster representative sequences using the KEGG/KAAS web-server [Moriya et al., 2007] (using singledirectional best hit EST mode against a eukaryote representative gene set, bit-score cut-off $\geq$40). The resulting KO (Kegg Orthology) lists were scaled by cluster size and filtered using MinPath [Ye and Doak, 2009] to get a minimal set of pathways. Hits for KEGG pathways K000230: Purine metabolism and K000240: Pyrimidine metabolism were summed and plotted against temperature for each environment.

GO

We then mapped all detected Pfam domains to their corresponding GO term(s) [Ashburner et al., 2000] using a custom Perl script and the mapping file Pfam2Go (http://www.geneontology.org/external2go/pfam2go). Then, for each possible pair of environments we performed a Fisher's exact test on each GO term. Enriched GO terms (For all three categories: Cellular component, biological process and molecular function) were identified using a Bonferroni corrected p-value $<$0.001 and used to create term clouds (One for each environment in the pairwise comparison). Lists of enriched GO terms were created - one for each environment, with the frequency of a GO term in the list determined by the absolute difference in the normalised abundance of the term between the two environments. Term clouds were then created using Worditout.com using direct colour blending from blue (low frequency) to red (high frequency).

3.3.5 Comparative analyses

Heatmaps

We created heatmaps to summarise both the PhymmBL taxonomic classifications and Gene Ontology totals using the Heatmap.2 function in R. Heatmaps represent frequency data using a matrix of colour coded cells (where the colours represent the frequency of the value). Either the rows of the matrix, the columns or both can be hierarchically clustered, grouping similar rows or columns in close proximity. For the taxonomy heatmap, only PhymmBL classified algal groups were used. For GO terms we only used biological process GO terms that were present over a certain abundance cutoff ($\geq$0.5% of hits in at least one data set). For the column dendrogram representing the overall similarity of the data sets, we read in the percentage of hits to each phyla or GO term as a table and used these to create a distance matrix using the value: 1 minus the Pearson correlation coefficient. The row dendrogram was created using the default settings of Heatmap.2 (euclidean distance). Both dendrograms were created using complete-linkage clustering, and the data cells were scaled and centred by column.

Multidimensional scaling

Next we performed multidimensional scaling for the most abundant phyla – the Bacillariophyta and Dinoflagellata – based on the set of Pfam domains detected for these organisms. Multidimensional scaling is a technique to place objects in n-dimensional space based on a correlation matrix. The proximity of objects to one-another in the plot reflects how similar they are. We extracted Pfam results for Bacillariophyta and Dinoflagellata sequences (PhymmBL phylum confidence score $\geq$0.9) and used them to create a 10x10 correlation matrix, with each cell of the matrix holding the correlation score (1 minus the Spearman correlation coefficient) between the Pfam totals of the two groups represented by the cell. The correlation matrix was converted to a euclidean distance matrix and non-metric multidimensional scaling performed using the IsoMDS function (Kruskall's) of the MASS package in R.

Canonical correspondence analysis

We performed a canonical correspondence analysis (CCA) using the VEGAN package in R. CCA is a two-stage technique consisting of ordination and environmental gradient identification fitting [Ter Braak, 1986], and is commonly used for analysing the effect of environmental variables on community composition. We treated the transpose of the normalised Pfam count tables as our species data and created a second table of environmental factors such as temperature, salinity, latitude, longitude and nutrient levels. Where environmental data was unavailable we used the World Ocean Atlas (http://www.nodc.noaa.gov/OC5/SELECT/woaselect/woaselect.html) for nutrient levels, taking the annual surface mean values. For light levels we used the Pangaea information system website (http://www.pangaea.de) to find insitu PAR (Photosynthetically Active Radiation) readings over a depth gradient for environments analogous to our samples. By plotting PAR against depth and fitting an exponential regression line we could extract the equation for the PAR - depth relationship and plug in our depth measurement to get an estimated PAR. The data sets we used were: ANT (Nicolaus, M et al. (2012): Downward spectral solar irradiance as measured in different depths under sea ice (transmitted irradiance) at sea ice station PS78/267-1.doi:10.1594/PANGAEA.786857);

EPAC (Eldin, Gerard; Rodier, Martine; Dupouy, D (2004): Physical oceanography at CTD station FLUPAC_119. doi:10.1594/PANGAEA.186766); ARC and NATL (Fossa, Jan Helge; Kutti, Tina; Bergstad, Odd Aksel; Knutsen, Tor; Svellingen, Ingvald; Wangensten, Jarle; Johannessen, Reidar; Steinsland, Asgeir (2011): Physical oceanography during R/V H. Mosby cruise IMR-2009615. Institute of Marine Research, Bergen, doi:10.1594/PANGAEA.756308); NPAC (NPAC: Whitney, Frank (2002): Physical oceanography at station IOS_97-11_CTD045.doi:10.1594/PANGAEA.79563).

Where there were multiple samples, such as in the ANT and EPAC samples, we took the mean values. All environmental data were transformed to a log_2 scale and an offset added to temperature values to make them positive. See Appendix A.4 for a table of environmental conditions. To highlight specific proteins for nitrate reductases, fucoxanthin chlorophyll binding proteins (FCPs), ribosomal proteins and silicon transporters we took one gene of each type from 3 diatoms: *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* and *Cylindrotheca fusiformis*. Each gene was compared to Pfam-A (gathering threshold cutoff) and the detected domains used to represent that gene.

Comparisons with species specific transcriptome

As ~60% of sequences in the NPAC sample had taxonomic affiliations with *Thalassiosira pseudonana* we chose this data set to perform a comparison with expression data from a *T. pseudonana* genome-wide microarray experiment [Mock et al., 2008]. First we compiled a spreadsheet of differentially expressed (log_2 fold change $\leq$1, p-value <0.05) *T. pseudonana* genes and expression values under low temperature (4°C), and silicate, nitrate, iron and CO2 limitation. Columns were added to each gene for GO, KEGG, KOG (annotations taken from JGI `http://genome.jgipsf.org/Thaps3/Thaps3.download.ftp.html`)) and Pfam annotations (performed ourselves by searching against Pfam-A and using the default gathering threshold cutoffs). Sequences from the NPAC sample classified as Bacillariophyta (PhymmBL phylum confidence score ≥0.9) were extracted and BLASTed against the JGI *T. pseudonana* gene models (BLASTX, e-value ≤1e-5, using soft masking, requiring ≥50% coverage of the query and ≥75% overall identity and taking the single best hit). Finally we added the number of total matches to each differentially expressed gene to the table.

3.4 Results

3.4.1 Sequence processing

The combined 454 sequencing runs produced 5 data sets giving us a total of 2,075,984 million raw sequences. After quality filtering we were left with 1,533,513 sequences, that is ∼74% of the original total. For four of the data sets, between 18% and 28% of raw sequences were filtered out completely, however for the ANT data set ∼44% of sequences were removed. This was caused by a combination of low complexity sequences and a high proportion of putative sequencing artifacts i.e. artificial duplicates. The EPAC and NPAC data sets also contained a considerably higher number of artifacts than the ARC and NATL data sets. We do not know whether this reflects genuine sequencing artifacts or is due to natural duplicates caused by a highly dominant species or dominant transcript. However the low number of clusters produced from these data sets (see table 3.1) seems to reflect a reduced level of sequence diversity.

The NATL and ARC data sets displayed some other clear differences to

	ANT	ARC	EPAC	NATL	NPAC
#Raw Reads	391,614	514,223	342,252	513,985	313,910
Avg Length (bp)	168.1	278.1	158.9	310.7	258
Total Size (Mb)	65.83	143.03	54.4	159.67	81
Potential Artifacts[1]	49,093	3,175	21,942	5,172	14,172
Putative rRNA[2]	3,595	38,651	1,324	68,009	1,254
#Filtered Reads	220,844	421,107	246,534	394,187	250,841
Avg Length (bp)	209.3	252.1	161	285.6	268.2
Total Size (Mb)	46.22	106.18	39.69	112.58	67.26
GC%	43.43	43.82	47.3	43.99	44.44
#Clusters[3]	29,840	254,423	119,783	252,031	76,564

Table 3.1: Summary of 454 sequence data for Antarctic (ANT), Arctic (ARC), Equatorial Pacific (EPAC), North Atlantic (NATL), and North Pacific (NPAC) metatranscriptomes. 1: Only exact duplicates were removed: CD-HIT-est clustering at 100% identity requiring 100% coverage of both sequences. 2: BLASTN against Silva SSU & LSU database Best hit, no complexity filtering, bit-score cutoff ≥50. 3: CD-HIT-est clustering ≥95% overall identity, requiring ≥50% coverage cluster representative. ANT, EPAC and NPAC sequenced with GS-FLX, ARC and NATL sequenced with GS-FLX Titanium.

the other data sets; the sole use of the Titanium platform (the EPAC, NPAC and ANT data sets were produced using a combination of FLX and Titanium sequences) was reflected in the longer average read length (see table 3.1). Also, they contained a much higher proportion of putative rRNA sequences: 7.5% and 13% of raw reads in ARC and NATL respectively compared to <1% in the other three data sets. This is likely due to differences in sample preparation; the employment of 2 rounds of mRNA purification used to prepare the ANT, EPAC and NPAC samples having removed virtually all non-mRNA.

Exploratory clustering of sequences from all 5 samples showed that the majority of sequences were environment specific, from 82% in NPAC to as much as ∼97% of sequences in the ANT and EPAC data sets (see figure 3.3). Only a handful of sequences (552) were contained in clusters containing sequences from all five environments.

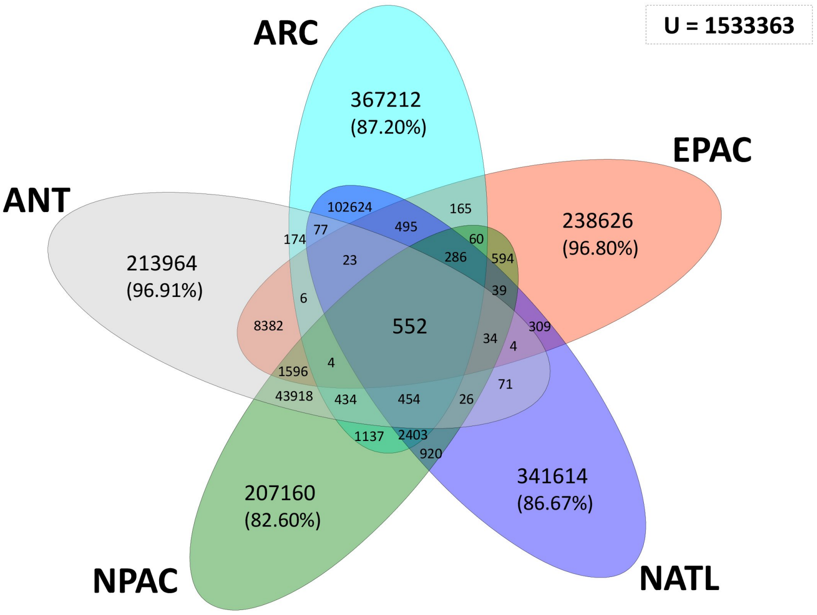


Figure 3.3: Sequence-distribution Venn diagram for pooled sequences clustering based on CD-HIT (longest open reading frames clustered using 90% similarity and 50% overlap of sequences).

The largest overlap was between the ARC and NATL data sets, 102,624 sequences fell into clusters shared by these two samples. This is perhaps unsurprising considering the close proximity of the two sampling sites. The second largest overlap was between the NPAC and ANT data sets (43,918 sequences). This cannot be explained by geographic location and may be explained by similarities in population composition or transcriptional behaviour.

3.4.2 Rarefaction curves

The rarefaction curves based on Pfam protein domains showed a levelling-off for all five samples with all sequences included (see figure 3.4). Compared to the Chao-1 estimates, our samples contained between 70% and 85% of the predicted domain content. We produced rarefaction curves based on detected protein domains rather than taxa; the limited number of reference genomes used in this analysis would lead to the curves plateauing too soon and producing misleading results.

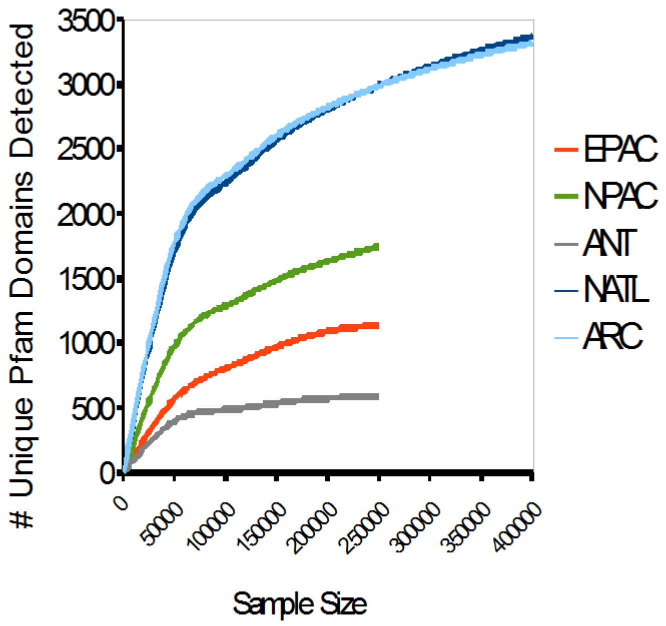


Figure 3.4: Pfam protein domain rarefaction curves for Equatorial Pacific (EPAC), North Pacific (NPAC), Antarctic (ANT), North Altantic (NATL) and Arctic (ARC) metatranscriptomes. (Chao-1 estimator of species richness using 50,000 sequence increments. http://www.biology.ualberta.ca/jbrzusto/rarefact.php).

3.4.3 Taxonomic affiliations

A relatively low proportion of sequences returned matches to organisms in the PhymmBL reference database. Using a confidence score cut-off of ≥0.9 at the phylum level, the proportion of sequences assigned a reliable taxonomic affiliation ranged from as low as 3% in EPAC to 33% in ANT, with around

8% of sequences from the other 3 data sets returning matches to our reference database.

Although all samples contained a small proportion of matches to out-groups, the EPAC sample contained a higher percentage of matches to bacterial sequences. However, the number of overall matches was generally very low - the most abundant transcripts in all 5 data sets were from eukaryotic phytoplankton. Hierarchical clustering of species abundance, separated the samples into two clusters, the diatom dominated NPAC and polar samples (ANT and ARC) and the dinoflagellate/ciliophora dominated open ocean samples EPAC and NATL (see figure 3.5).

Diatoms are known to dominate phytoplankton communities in coastal upwelling systems such as Puget Sound and sea ice [Armbrust, 2009]. This was reflected at the species level taxonomic affiliations: 61% of assignable sequences from NPAC were most similar to *Thalassiosira pseudonana* (this is consistent with microscopic observations, see Appendix A.5), and 86% of assignable sequences from ANT were most similar to the polar diatom *Fragilariopsis cylindrus*.

3.4.4 Functional annotation

As with other metatranscriptomics projects (see section 2.2) between 16% and 35% of sequences returned matches to Pfam domains. However, as for the taxonomic affiliations, the EPAC data set returned significantly fewer matches than the other 4 samples (around 5.7%). The ARC and NATL samples returned the most diverse range of domains with ∼3,000 unique domains identified, the ANT sample returned the fewest (583 unique domains). This probably reflects the dominance of a single species (see above) in a niche environment (low temperature, high salinity brine pockets in sea ice) performing a specialised set

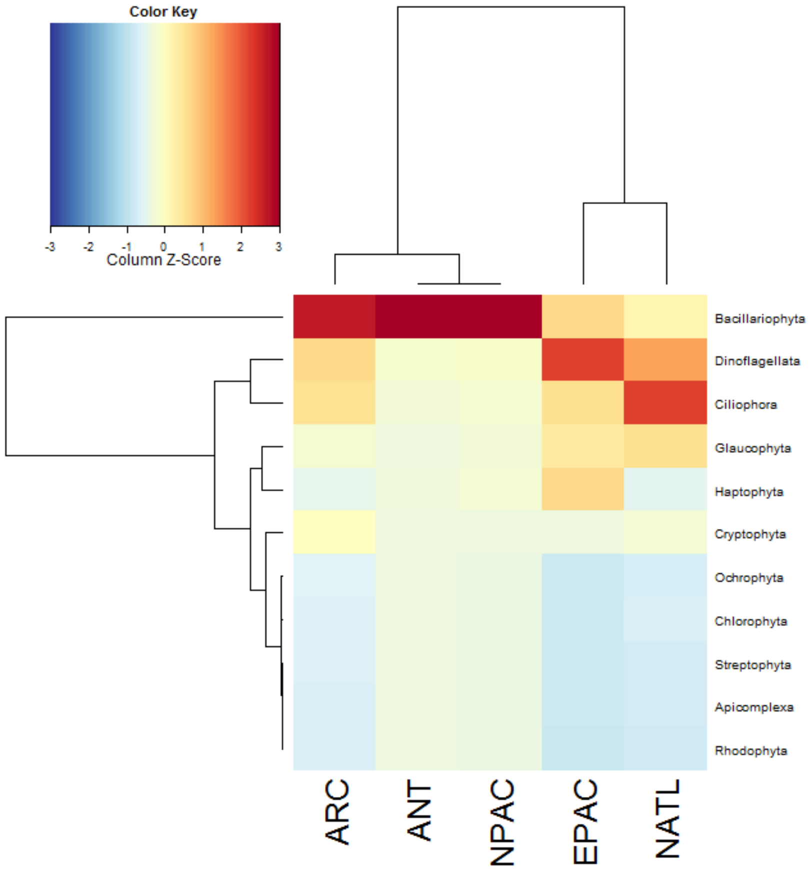


Figure 3.5: Heatmap summary of PhymmBL-classified (confidence score $\geq$0.9) algal sequence abundances. Complete-linkage clustering was employed for both row and column dendrograms. Column dendrogram created from a distance matrix of 1-Pearson correlation coefficients; row dendrogram created using default dist function of Heatmap.2 (euclidean distance matrix). Cell values were scaled by column: z-score represents the original value minus the column mean and divided by the standard deviation.

of metabolic functions.

The most abundant domains were components of the ribosome, a major component in the biosynthesis of proteins. These comprised between 41% and 9% (ribosomes were the second most abundant domain in EPAC, after the bac_rhodopsin) of all identified domains. A heatmap of biological process GO terms showed that translation was the most abundant process (see figure 3.6).

It also indicated an apparent gradient in the abundance of sequences associated with translation, with the lowest abundance in the tropical sample (EPAC) increasing in the temperate samples (NATL and NPAC) and the highest values in the two polar samples (ANT and ARC). A strong correlation was detected between the normalised abundance of sequences associated with the GO term for translation and the in-situ temperature of the sampling sites ($R^2 = 0.8$).

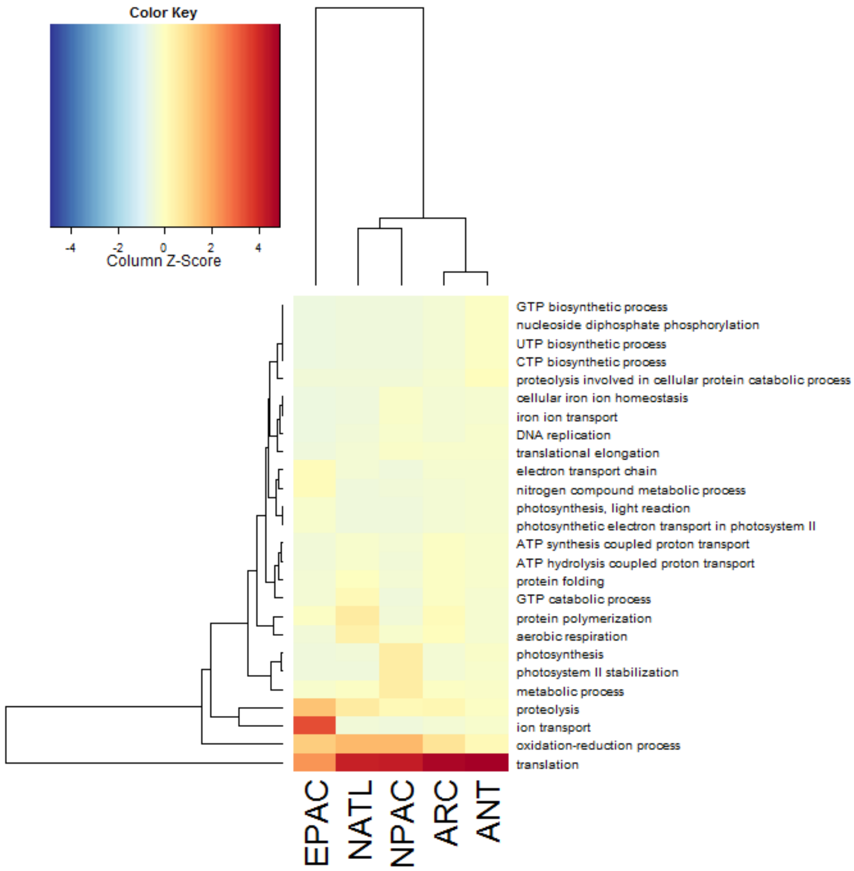


Figure 3.6: Heatmap summary of biological process GO terms. Complete-linkage clustering was employed for both row and column dendrograms. Column dendrogram created from a distance matrix of 1-Pearson correlation coefficients; row dendrogram created using default dist function of Heatmap.2 (euclidean distance matrix). Cell values were scaled by column: z-score represents the original value minus the column mean and divided by the standard deviation.

This was also reflected in our pairwise GO-term word clouds. When looking at significantly different GO terms (Bonferroni corrected p-value <0.001) it was clear that in lower temperature samples, GO terms associated with biosynthesis, such as 'translation', 'ribosome' were enriched. See figure 3.7 below for GO term cloud of significantly enriched GO terms in ANT compared to EPAC.

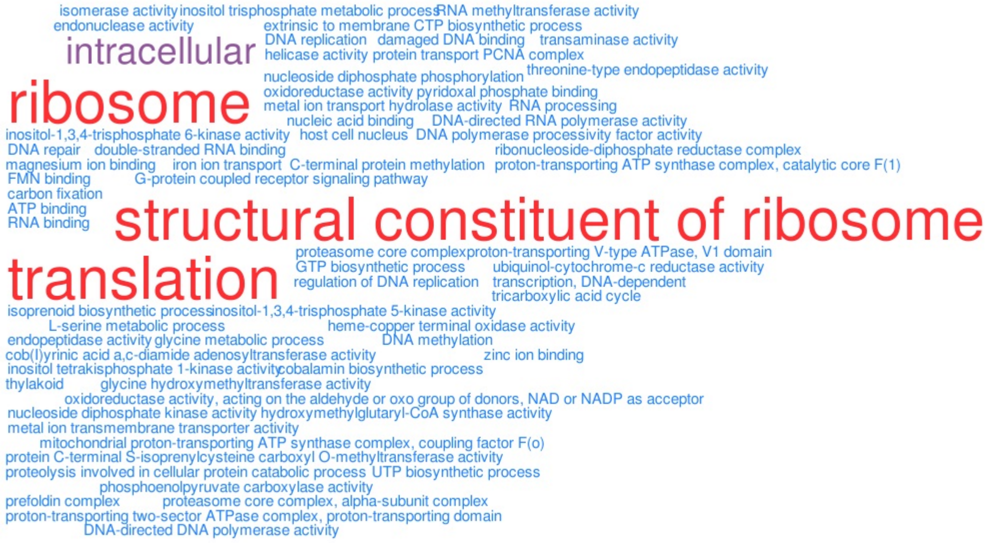


Figure 3.7: Statistically significantly (Bonferroni corrected Fisher's exact test p-value <0.001) enriched GO terms in ANT compared to EPAC. Term clouds created with http://www.worditout.com. Terms scaled by the absolute difference in the relative abundance of the enriched term and using direct colour blending from blue (low frequency) to red (high frequency).

3.4.5 Comparative analyses

Multidimensional scaling

As bacillariophyta and dinoflagellata were the most abundant taxa in our samples, we selected these groups for our MDS plot. The final ordination required four dimensions with a Kruskall's stress value of 0.0086 (measure of goodness of fit for the ordination, which ideally should be ≤ 0.1 [Manly, 2005]) after 50 iterations. In figure 3.8 the first two dimensions show a clear separation along the x-axis of bacillariophyta and dinoflagellata based on their transcriptional profiles reflecting the different transcriptional behaviour of these taxa. The bacillariophyta samples also appear to have been positioned along the y-axis (Dimension 2) according to the latitude of their sampling sites. In addition, the y-coordinates of the bacillariophyta samples had a very strong correlation with latitude ($R^2 = 0.99$).

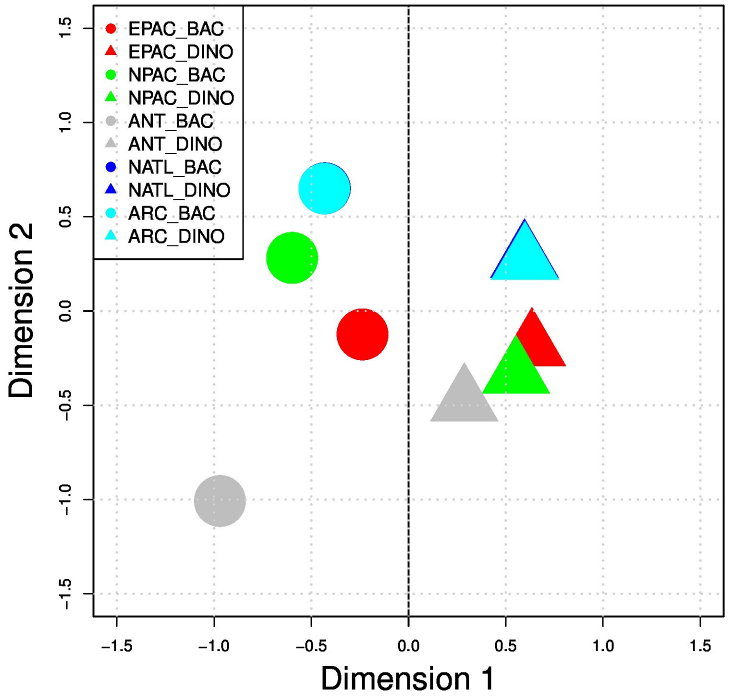


Figure 3.8: Non-metric multidimensional scaling (MDS) plot based on a distance matrix (1 minus Spearman correlation coefficient) of Pfam protein families from PhymmBL classified Bacillariophyta (BAC) and Dinoflagellata (DINO) sequences.

Canonical correspondence analysis

A multiple correlation plot of environmental factors with the abundance of transcripts in each environment associated with the GO term for translation identified temperature as the most strongly correlated factor ($R^2 = 0.81$) see Appendix A.6). Using a set of four environmental conditions (temperature, light, nitrate and phosphate) 100% of variability could be accounted for. Four dimensions were required to account for the variability in the data, the dimensions accounted for 37.31%, 31.84%, 26.5% and 4.3% of total variability respectively. Taken in isolation the four factors could account for 34.87% (phosphate), 31.71% (nitrate), 30.17% (light) and 28.32% temperature.

Plotting the first three dimensions shows that the nutrients nitrate and phosphate are mostly strongly associated with the first dimension (reflected in the direction and length of vector) and that light and temperature are most strongly associated with dimension 2 (see Appendix A.7 for CCA plots of dimensions 1 and 2 and 1 and 3). We highlighted all ribosomal proteins in the plots, and the plot of dimensions 2 and 3 shows that most of the variation in ribosomal proteins occurs along dimension 2, suggesting that the environmental factors light and temperature are strong influences on this variability, see figure 3.9 below.

Comparisons with species specific transcriptome

We extracted 14,926 sequences from the NPAC metatranscriptome identified as bacillariophyta-like and aligned these to *T. pseudonana* genes. A total of 10,713 sequences aligned within the thresholds used. Over 95% of these sequences matched to *T. pseudonana* genes that were up-regulated under low temperature.

Although Puget Sound, the source of the NPAC samples, is classed as

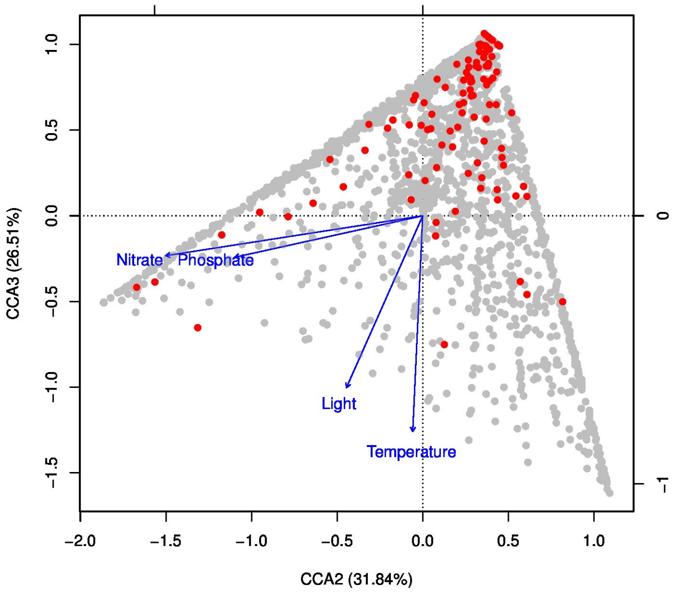


Figure 3.9: Canonical correspondence analysis (CCA) between protein family (Pfam) abundance and environmental conditions deduced from ocean samples in this study, red circles represent ribosomal transcripts.

a temperate system, the average surface temperature doesn't exceed 12°C, even in summer [Moore et al., 2008] (the temperature at the time of sampling of 12°C reflecting the annual maximum). Also, several strains of *T. pseudonana* have been shown to be growth limited by temperatures of 13.5°C [Ferguson et al., 1976]. So, it would seem that naturally occurring communities of *T. pseudonana*-like species were limited by the low temperature in Puget Sound.

3.4.6 Follow up experiments

To test this relationship between temperature and the transcription of ribosomal genes detected through *in-silico* analysis, laboratory experiments were performed on model diatom species under control conditions. The first of these was quantitative real time polymerase chain reaction (qPCR), a technique to measure the abundance of specific DNA molecules (or reverse transcribed mRNA). The experiment showed that for five different ribosomal genes of the diatom *Fragilariopsis cylindrus*, levels of mRNA were significantly increased (log_2 fold change ≥ 1, p <0.05) at -2°C compared with 10°C.

In order to determine whether this increase in the transcription of ribosomal genes led to an increase in the production of ribosomes, Western blots were performed for two model diatom species - *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*. Western blots are a technique used to detect and estimate the abundance of proteins. It uses gel electrophoresis to separate protein samples by molecular weight, the proteins are then stained by protein-specific antibodies, the thicker the band the more abundant the protein. The ribosomal protein S14 was used for both diatoms under nutrient replete conditions across a range of temperatures. Figure 3.10 clearly shows an increased presence of this protein for decreasing temperatures.

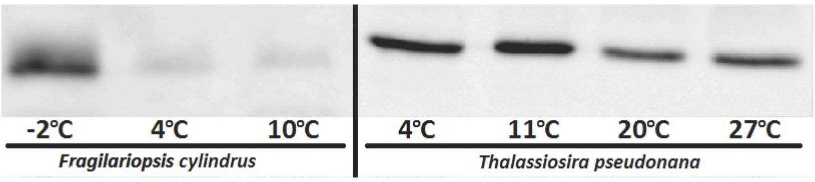


Figure 3.10: Western Blots using a commercial antipeptide against the eukaryotic ribosomal protein S14. Cultures of *Fragilariopsis cylindrus* and *Thalassiosira pseudonana* were cultivated at different temperatures under nutrient replete conditions.

The final follow up experiment was to test whether the rate of translation was affected by temperature. The Mock Lab performed a translation efficiency experiment using a transgenic strain of *Thalassiosira pseudonana*, with a particular gene (nitrate reductase) modified to express GFP (Green Fluoresence

Protein) when nitrate is added to the culture (i.e. when the nitrate reductase is translated). Measurements of the percentage increase of GFP produced over time were used to reflect translational efficiency. The results showed that the amount of GFP expressed increased at a faster rate (∼3 fold) for*Thalassiosira pseudonana* cultures at 20°C than at 11°C (See Appendix A.8). So, despite an increase in the abundance of ribosomes at lower temperatures, it seems that the rate of translational efficiency is reduced. This increase in the production of ribosomes may be a compensatory mechanism, as the rate of translation is reduced more ribosomes are required for biosynthesis.

3.5 Discussion

In this chapter we have described the analysis of metatranscriptome samples from eukaryotic phytoplankton as part of an integrative approach combining bioinformatics and molecular biology. The metatranscriptomics computational pipeline involved quality filtering of high-throughput sequence data, identifying the taxonomic composition of samples, determining the transcriptional profile and a series of comparative analyses. Multivariate statistical techniques such as hierarchical clustering, multidimensional scaling and canonical correspondence analysis were employed to identify similarities and differences between samples and to assess the influence of environmental conditions on transcriptional profiles. Not only is this the first large-scale, global eukaryotic phytoplankton metatranscriptome study, but it also led to new biological discoveries.

The main shortcomings of this project lie in the experimental design. Firstly, the project consists of just five sample sites. It is therefore impossible to say if these samples are truely reflective of the environment types they represent. The small number of samples may also weaken the significance of statistical methods employed and any differences between samples may be exaggerated. Secondly, there was no replication of sampling. It is common practice to produce replicate (usually triplicate) samples to help differentiate between biological variability and technical variability. With only a single sample it is difficult to determine whether transcripts expressed at high or low levels reflective genuine microbial behaviour or are simply an artifact of the particular sample. Thirdly, the samples were collected by different groups using slightly different approaches and these were sequenced by different facilities using a combination of different platforms. This could also lead to biases in the results.

The *in-silico* analysis identified phytoplankton species as the dominant taxa in all five samples and found that the majority of transcripts were associated with the biosynthesis of proteins. In addition, we identified a strong relationship between temperature and translation (a strong correlation between temperature and the abundance of transcripts identified as ribosomal - a key component in the biosynthesis of proteins). Follow-up molecular biology experiments showed that lower temperatures lead to increased levels of mRNA for ribosomal genes, and increased levels of ribosomal proteins. In addition, a translational efficiency experiment demonstrated that the rate of translation (for a specific gene) was significantly lower at lower temperatures.

We also collaborated with a group from the College of Life and Environmental Sciences from the University of Exeter. They produced both cell-level resource allocation models and global models of cellular nitrate and phosphate levels. The ratio of nitrate to phosphate in plankton is part of the Redfield ratio (the rate of carbon to nitrogen to phosphate atoms, traditionally 106:16:1). Recent worked revealed that the N:P component of the Redfield ratio is re-

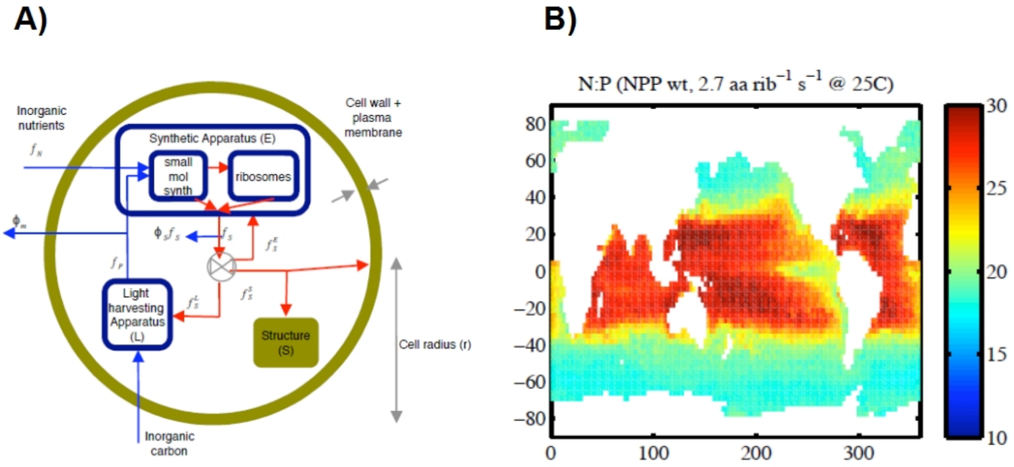


Figure 3.11: A) Cell model containing light-harvesting apparatus (L) (Chlorophyll, accessory pigments), biosynthesis (E) (small molecule biosynthesis (sm) and ribosomes (rib)) and cell structure (S) (including nutrient acquisition and assimilation). B) Modelled N:P ratios based on cell and ecosystems model.

lated to the ratio of nitrogen rich protein to phosphate rich ribosomal RNA (rRNA) [Loladze and Elser, 2011]. The global models predict that N:P ratios will be lower in cold regions where resource allocation emphasises biosynthesis see figure 3.11-B, whereas in warmer regions N:P will be higher due to emphasis on photosynthesis. It seems likely therefore that increased water temperature due to anthropogenic global warming could cause changes in the fundamental chemistry of plankton communities and the ocean itself.

Chapter 4

Assessment of sequence processing methods on simulated metatranscriptome data

This chapter is adapted from ‘Assessment of sequence processing methods on simulated metatranscriptome data ’, A. Toseland, S. Moxon, T. Mock, V. Moulton, in preparation.

4.1 Summary

This chapter describes the assessment of sequence clustering on simulated metatranscriptome sequence data in order to determine an optimal parameter set. We developed an approach to simulate metatranscriptome data and assessed a widely used clustering application over an extensive range of parameters. We also compared this approach with an alternative sequence processing method - sequence assembly.

4.2 Background

Metatranscriptome data can contain a high degree of redundancy, that is, multiple identical or nearly identical sequences. In an investigation into the proportion of artificial and natural duplicates in pyrosequenced metatranscriptome data, Niu et al. reported that as much as 60% of all sequences in an early metatranscriptome data set were likely natural duplicates [Niu et al., 2010]. Generally only a small proportion (∼15-35%) of metatranscriptome sequences have homologs in reference databases [Frias-Lopez et al., 2008], [Poretsky et al., 2009], [Gifford et al., 2010], [Qi et al., 2011]. Therefore, some form of data reduction strategy is beneficial before running computationally intensive homology searches.

Two approaches that are often employed to reduce redundancy in metatranscriptome sequence data sets are to assemble sequences into contiguous fragments or to separate the data set into clusters of sequences sharing a defined degree of similarity. Metatranscriptome data shares some properties of metagenomic data sets, namely that the sequence data derives from a diverse range of organisms and that the coverage of each will be highly variable.

This can lead to fragmented assemblies containing potentially chimeric contigs [Kunin et al., 2008] and recommendations have been made that highly diverse data sets should not be assembled [Mavromatis et al., 2007] and that clustering be used as an alternative data reduction technique [Thomas et al., 2012]. Some authors of metatranscriptome projects assemble [Ogura et al., 2011], some cluster as a data reduction strategy [Gilbert et al., 2008], [Poretsky et al., 2009], [Rinta-Kanto et al., 2012] or as a means of removing potential artificial duplicates [Stewart et al., 2011] and some do neither [Frias-Lopez et al., 2008] [Urich et al., 2008].

In this chapter, we shall investigate popular data reduction tools and assess their performance for metatranscriptome data in terms of the accuracy of resulting protein annotations. To do this we will also describe a way to simulate such data sets. Several approaches have previously been described to simulate metagenomic data sets [Mavromatis et al., 2007], [Pignatelli and Moya, 2011], [Mende et al., 2012] for benchmarking assembly and gene annotation tools. However, to date no similar methods have been developed to simulate metatranscriptomic data for similar purposes.

4.3 Materials and methods

4.3.1 Simulated metatranscriptomes

We created three population profiles to represent low, medium and high diversity bacterial communities (referred to as LD, MD and HD respectively from here on). These were recreated as closely as possible from the organism lists and genome coverage levels used in the simulated metagenome study by Pignatelli et al. [Pignatelli and Moya, 2011]. All 3 populations contained the same

112 organisms used in the Pignatelli simulations but are present in different quantities. The genome coverage values were scaled to create discrete organism abundances to give a total population size of approximately 1,000 for each sample. The low diversity population was dominated by a single taxa and represents a niche environment such as the acid mine drainage metagenome sampled by Tyson et al. [Tyson et al., 2004]. The medium diversity set contains a small number of dominant taxa, and finally the high diversity sample contains no dominant taxa, all organisms are present in roughly equal proportions (See Appendix B.1 for list of organisms used).

For each diversity level, we generated a set of species-specific transcript expression profiles. For each of the 112 species in the samples, we generated a Pareto-like, power law distribution ($P(k) \propto \mathsf{k}^{-r}$) [Ueda et al., 2004], modelling the probability of a gene having the level of expression k and the exponent r is directly related to the rate of mRNA decay [Nacher and Akutsu, 2006]. This distribution has been empirically demonstrated (based on genome-wide microarray data) to apply to gene expression from a range of model organisms such as bacteria (*E. coli*), yeast (*S. cerevisiae*), plant (*A. thaliana*), insect (*D. melanogaster*) and mammal (*M. musculus* and *H. sapiens*) [Ueda et al., 2004]. For each species we used J. Cristobal Vera's transcript simulator (`http:/personal.psu.edu/jcv128/software.html`) to produce an expression profile using an r exponent of 1.69 (*E. coli* value as shown by [Ueda et al., 2004]), each gene could take an expression value between 1 and 1,000 within a Pareto power law distribution, reflecting the number of transcript copies present in the cell, which is then scaled up by the total abundance of the organism in the sample.

Next, we downloaded the gene sequences (nucleotide and amino acid sequences) for all 112 species from the Joint Genome Institutes Integrated Microbial Genomes database (JGI-IMG) [Markowitz et al., 2006]. For each diversity level we first created a test data set for comparative purposes using the JGI-IMG manually curated, error-free gene models in translated amino acid format (these data sets are referred to with the suffix 'AA'). We sampled each gene a number of times equal to the copy number from the appropriate expression profile using a random start location and an average sequence length of 100 amino acids (assuming a 454 nucleotide sequence of 200-400 bp translated in full). All sampled fragments were added to a sequence file representing the transcript pool.

To introduce more realism, the second sample sets used the JGI-IMG nucleotide gene models (referred to with the suffix 'NT'). We again randomly sampled fragments from each gene a number of times equal to the copy number, we used a random start location, a minimum length of 6 bases to reflect random hexamer primers and a maximum length of 400 bases. We then ran these fragments through 454sim [Lysholm et al., 2011] using the GS-FLX error models to introduce realistic sequence errors and translated the resulting sequence into their longest open reading frames. For both the AA and NT data sets we then randomly sampled 250,000 sequences without replacement from each transcript pool. Although the 454 GS-FLX platform can produce ∼400,000 sequences per run [Shendure and Ji, 2008], after quality filtering and removal of rRNA for example, there are often less than 300,000 sequences remaining (see table 3.1 for example).

4.3.2 Clustering

Sequence clustering programs divide a set of nucleotide or amino acid sequences into groups sharing a specified degree of similarity. Each cluster of sequences is represented by the longest sequence in the cluster. These representative sequences can then be annotated by homology searches against databases such as NCBI-nr [Pruitt et al., 2007], Pfam [Finn et al., 2010], COG [Tatusov et al., 2003] and KEGG [Ogata et al., 1999]. The resulting annotations can then be transferred to the shorter, member sequences of the cluster, some of which may contain insufficient protein domain regions to allow classification. If the clustering parameters used are too loose, sequences with unrelated function may be grouped together and transferred annotations may be false positives, but, if the parameters are too stringent, then little benefit will be gained in terms of data reduction.

We performed all clustering using CD-HIT [Li and Godzik, 2006], a popular clustering application due to its a high-speed, short word filtering algorithm and range of utility programs. For the nucleotide simulated metatranscriptome data, the sequences were translated into a set of longest open reading frames. We chose to cluster amino acid sequences rather than nucleotide sequences as synonymous codons could lead to nucleotide sequences that translate into amino acid sequences with high similarity being assigned to separate clusters. The sequences were clustered over a range of identity parameters. A nested loop was used to increment overall sequence similarity (C) from 40% to 100% (in 20% increments), and then percentage coverage of the cluster representative (aL) and cluster members (aS) increasing in 25% increments from 0 to 100%.

4.3.3 Assembly

The simulated 454 nucleotide data sets and the two real metatranscriptomes were assembled using MIRA [Chevreux et al., 2004], in de-novo, accurate, est mode, with non-uniform read depth, and all other parameters as default. MIRA features an extensive range of configurable parameters, making it a popular choice for assembling complex data sets. Both the contigs and debris (unassembled sequences) were translated into their longest open reading frames as above and Pfam homology searches were performed on all contig and debris sequences.

The final stage was a two-step processing stage of first assembling the nucleotide sequences as above, combining the translated contigs and debris and clustering them using the same range of parameters as in the clustering experiments described above.

4.3.4 Sensitivity, specificity and accuracy

The sensitivity, specificity and accuracy of annotation was assessed for all reads (no clustering or assembly); for the transferred annotation of cluster representatives to cluster members; for assembled contigs and singletons; and finally, for clustered assemblies. Sequences were compared to the Pfam-A database using pfam_scan.pl and detected domains compared to the Pfam domains of the gene region the sequence originated from.

For each clustering parameter set we assessed the representative sequence by comparing the set of domains detected in the sequence itself with the annotation of the region of its origin. Each cluster member sequence was then compared to the cluster representative by comparing the representative sequence annotation with the region of origin of the cluster member. We classified simulated metatranscriptome sequences as containing a domain if the origin of the sequence overlapped a domain region by one or more bases.

For example, taking a cluster with representative sequence R and one member sequence M and the origin of the member O:

- If the representative sequence returns no domains ($\{R\} = 0$) and the region of origin of the member contains no partial or complete domains, ($\{M\} = 0$) we return 2 true negatives, one for the representative sequence and one for the member.

- If the representative sequence returns no domains ($\{R\} = 0$), but the region of origin of the member contains n partial or complete domains, ($\{M\} = n$) we return n false negatives.

- If both the representative and the member sequence contain domains we define (i) true positives as the intersection between the set of domains present in the representative and the set of domains contained in the origin of the member ($\{R\} \cap \{O\}$) (ii) false positives as the domains contained in the representative but not the origin of the member ($\{R\} - \{O\}$) and (iii) false negatives as domains contained in the member but not the representative ($\{M\} - \{R\}$).

The resulting counts of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are then used to calculate sensitivity (TP / TP + FN), specificity (TN / TN + FP) and accuracy (TP + TN / TP + FP + TN + FN) for each clustering parameter set.

4.3.5 Real metatranscriptomes

In addition to the simulated data sets, we also assessed the different sequence processing methods on two real prokaryotic 454 sequenced metatranscriptomes. We downloaded the 110m metatranscriptome (OMZ) from [Stewart et al., 2012] from the sequence read archive and the present-day mid bloom metatranscriptome (GIL) from [Gilbert et al., 2008] from CAMERA. The data sets were chosen to reflect real metatranscriptomes sequenced by 454 pyrosequencing, of a representative size in terms of the number of sequences and containing a low proportion of rRNA. The Gilbert data set was dominated by a couple of bacterial classes and reflects a medium diversity sample. In the OMZ data set, there is no clearly dominant taxa, and this data is somewhere between a medium and high diversity sample (see Appendix B.3). Both data sets were filtered to remove rRNA by BLAST searches against the Silva database (Both SSU and LSU, filtered using parameters in [Mou et al., 2008]). To remove potential sequencing artifacts, we also clustered each data set using CD-HIT (requiring 100% sequence similarity over 100% of the sequence length). The filtered sequences were then translated into their longest ORFs and clustered in the same way as for the simulations above. Although we cannot assess the resulting annotations in terms of accuracy (as we do not know the genuine domain content), these data sets are useful to assess the impact of the methods employed on real data.

4.4 Results

In summary, we have created 6 simulated metatranscriptomes. Three data sets representing three different levels of taxonomic diversity (low, medium and high) were sampled directly from error-free, correctly translated amino acid gene models. Three more data sets (also representing low, medium and high diversity populations) were sampled from the original nucleotide sequences of the same genes. These were then run through a 454 simulator to reflect sequencing errors and finally translated into longest open reading frames. We then assessed the accuracy of protein domain annotation against the Pfam-A database of each dataset on all unprocessed sequences, on clustered sequences, assembled sequences and finally by first assembling and then clustering the sequences. We also subjected two published metatranscriptome data sets to the same processing methods to provide an indication of how the methods work on real data.

4.4.1 Data sets

Each of the three simulated data sets consisted of 250,000 sequences with an average length of ∼200 bp (see table 4.1). Although 454 sequencing can produce longer sequences on average (around 250 bp) [Shendure and Ji, 2008] this is consistent with real 454 metatranscriptome data (see OMZ and GIL in table 4.1 and ANT, EPAC and NPAC data sets in table 3.1). This likely reflects the fact that we are sampling (using random start location due to random priming) from short fragments of mRNA (the average length of all JGI-IMG transcripts used for the simulations was <1,000 bp) rather than long stretches of genomic DNA and are therefore not guaranteed full length fragments.

We downloaded 557,762 and 162,871 sequences for the OMZ and GIL metatranscriptomes respectively. Our filtering steps identified ∼33% of sequences as putative rRNA for the OMZ data set and ∼3% for the GIL data set (see table 4.1). This differs by about 2-3% from the original publications. This is likely due to slight differences in the filtering parameters used. We were left with 313,147 and 153,487 non rRNA sequences for the OMZ and GIL samples respectively with an average length of over 200 bp.

Data Set	#Reads	#Nucleotides [Mb]	Avg. Length
LD	250000	50.4	201.5
MD	250000	49.7	198.6
HD	250000	51.1	204.3
GIL	153487	35.61	232
OMZ	313147	64.67	206.5

Table 4.1: Summary of 454 data sets. Number of reads, total number of nucleotides in Megabases and average sequence length for low diversity (LD), medium diversity (MD) and high diversity simulated metatranscriptomes, and for Gilbert (GIL) [Gilbert et al., 2008] and Stewart Oxygen Minimum Zone (OMZ) [Stewart et al., 2012] real metatranscriptomes.

4.4.2 Annotation of unprocessed sequences

Our next step was to assess the accuracy of Pfam annotations for all unprocessed sequences. This provides a baseline to compare subsequent results. For both the nucleotide (NT) and amino acid (AA) simulated data sets we performed homology searches against the Pfam-A database. Any detected domains were compared with the Pfam annotations for the origins of the simulated sequence (the region of the original JGI-IMG sequence our simulated read was sampled from) and assessed as in section 4.3.4.

AA data sets

For the AA simulated data sets, around 140,000 domains were detected in each diversity level (see table 4.2). Sensitivity for all 3 was 72%, specificity ranged between 78 and 89% and accuracy was between 74 and 76%.

Data Set	TP	FP	TN	FN	SENS	SPEC	ACC
LD-AA	145,895	8,616	57,544	57,306	0.72	0.87	0.76
MD-AA	143,428	7,858	60,920	56,277	0.72	0.89	0.76
HD-AA	138,367	16,369	59,337	54,433	0.72	0.78	0.74

Table 4.2: Assessment of Pfam annotations on all amino acid sequences for low diversity (LD), medium diversity (MD) and high diversity (HD) simulated metatranscriptomes.

NT data sets

For all three NT data sets around 30,000 domains were identified. While specificity was generally high (95-98%), sensitivity was significantly lower than for the AA simulations – around 16-17% (see table 4.3) and overall accuracy was around 40%. The difference in the number of domains detected between the AA simulations and the NT simulations is likely due to several factors. With the AA simulations, the sequences were taken from manually curated, error-free, correctly translated sequences. However, the NT simulations contain sequencing errors and are not guaranteed to be 'in-frame' (i.e. the first nucleotide of the sequence may not be the first position of a codon) and translation errors may occur. Lastly, the sequence length of the AA sequences (average 100 amino acids) may be overly optimistic, it is unlikely that a nucleotide sequence will translate in it's entirety to amino acids and it may be that only partial translations are possible.

Data Set	TP	FP	TN	FN	SENS	SPEC	ACC
LD-NT	30,406	1,961	76,350	159,548	0.16	0.97	0.40
MD-NT	30,112	1,917	79,683	156,166	0.16	0.98	0.41
HD-NT	31,770	4,376	81,860	149,787	0.17	0.95	0.42

Table 4.3: Assessment of Pfam annotations on all translated nucleotide sequences for low diversity (LD), medium diversity (MD) and high diversity (HD) simulated metatranscriptomes.

4.4.3 Clustered sequences

AA data sets

For the first data set, sampled directly from the JGI-IMG amino acid gene models, clustering produced a significant increase in the number of protein domains detected. The best performing parameter set, in terms of the largest positive difference between the increase in true positives and the increase in false positives (compared to annotating all sequences individually) was an overall similarity threshold of ≥40%, and requiring ≥25% coverage of the cluster representative and between ≥50-75% coverage of cluster member sequences. Increases in true positive detection of 5.88%, 6.49% and 6.45% were achieved for the LD, MD and HD data sets respectively (see figure 4.1).

NT data sets

For the translated nucleotide sequences, clustering at the lowest overall percentage similarity (40%) produced the poorest results. At this identity threshold, the number of additional false positives exceeded the number of additional true positives detected (see figure 4.1). As the overall percentage similarity threshold increased, both the number of true and false positives detected decreased. Also, as the required sequence coverage threshold increased the increase in true and false positive detection decreased.

The best performing parameter sets were ≥60% overall similarity, ≥0% coverage of the cluster representative and 0-50% minimum coverage of cluster members. These parameters produced increases in true positives of 5.82%, 7.48% and 4.79% for the LD, MD and HD data sets respectively. These increases came at the cost of a small increase in false positives, (see fig-

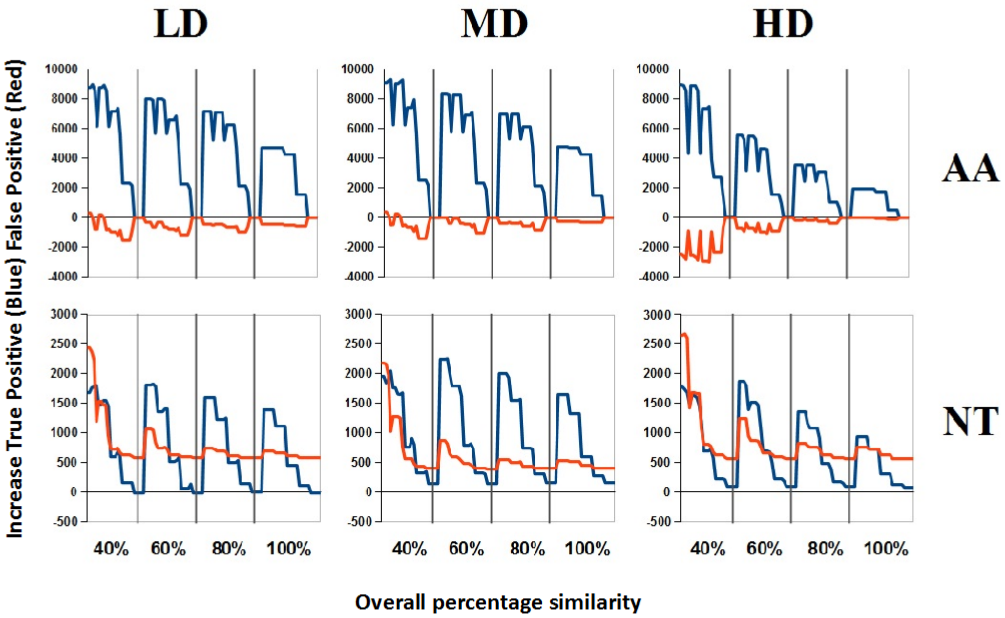


Figure 4.1: Increase in true positives (Blue line) and false positives (Red line) for annotated clustered sequences compared to annotation of all individual sequences for low diversity (LD), medium diversity (MD) and high diversity (HD) simulated metatranscriptomes. Top row - amino acid sequences directly sampled from JGI gene models. Bottom row - translated 454sim nucleotide sequences. X-axis represents clustering parameters, similarity in 20% increments. Within each 20% section are the results for clustering with varying coverage (0-100%) of the cluster representative and cluster members.

ure 4.1). Although sensitivity increased in all cases, the overall accuracy decreased slightly, due to a small increase in false positives, causing decreased specificity.

4.4.4 Assemblies

For the simulated nucleotide data sets, the assemblies incorporated around half of all sequences into contigs, with the exception of the high diversity data set (∼30% of all sequences). The average contig lengths were 298.6, 298.3 and 257.3 base pairs for LD-NT, MD-NT and HD-NT, respectively (see Appendix B.2 for assembly statistics). For the two real metatranscriptomes GIL and OMZ, 47.62% and 53.88% of sequences assembled into contigs with an average length of 415.5 and 244.4 respectively.

For both LD-NT and MD-NT data sets, the contigs alone produced more domains than both the baseline all sequence annotation and the best performing clustering parameter set. For the HD data set the combined annotations of the contigs and debris also exceeded the baseline and optimal clustering. The assemblies (combined contigs and debris) produced large increases in sensitivity (LD 13%, MD 14% and HD 4%), at a cost of decreased specificity (LD -4% ,MD -4% and HD 2%). However, overall accuracy increased by +7%, +8% and +2% for LD, MD and HD respectively.

Finally, the two stage approach of assembling sequences and then clustering the combined contigs and debris produced the highest overall increases in true positive domain detection. With the optimal clustering parameters (similarity of ≥60% and requiring ≥0% coverage of the representative and between 0-50% minimum coverage of cluster members) sensitivity was increased by (LD +14%, MD +15% and HD +4%). However, the slight increase in false positives decreased specificity (LD -5%, MD -7% and HD -4%) and overall accuracy increased in the LD and MD sets by 6% (see table 4.4 for full summary of sensitivity, specificity and accuracy for all 454 experiments).

The overall picture is that clustering produces a small increase in the number of true positives detected, however, overall accuracy is decreased slightly due to an increase in the number of false positives. Assembly produces a large increase in true positive detection, far outweighing the additional false positives introduced. Clustered assemblies produce the most true positives but again at the cost of a slight reduction in overall accuracy (see figure 4.2).

The low and medium diversity simulations produced the largest increases in true positive detection, and the overall results looked very similar (see figure 4.2). It would appear that samples containing one or more dominant species will tend to assemble well and these longer reads will allow for more domains to be classified more accurately. For the high diversity simulation, the different processing methods made relatively little difference. With no dominant species, it appears that little is to be gained from sequence assembly.

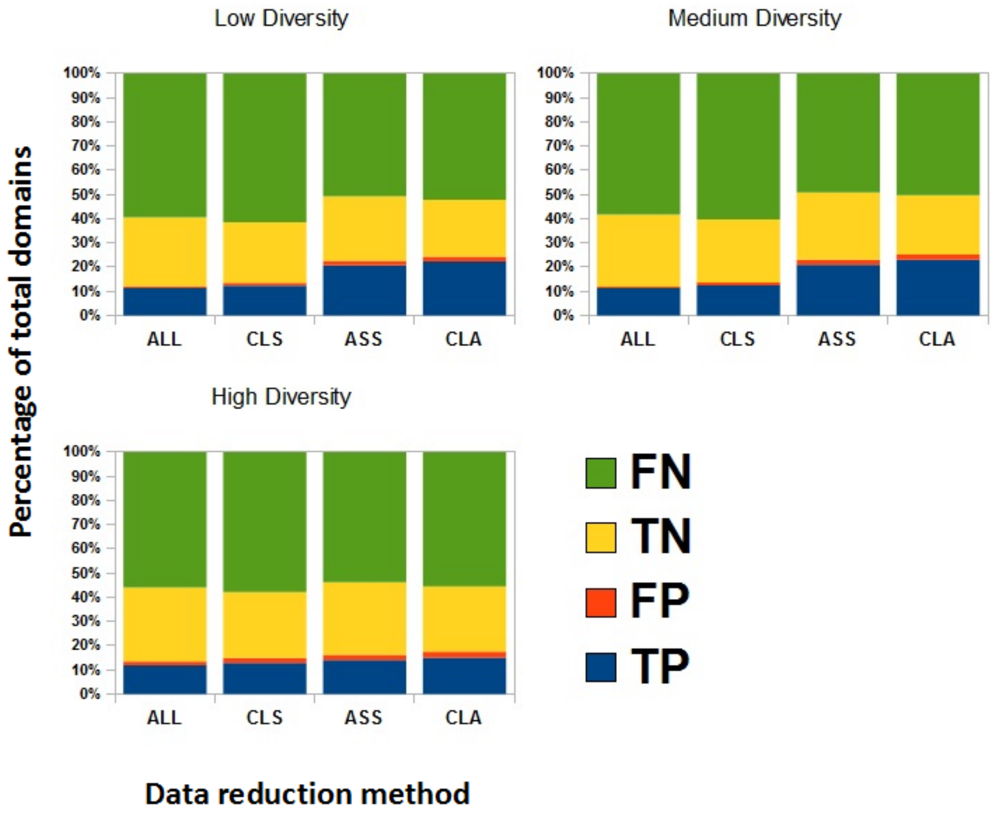


Figure 4.2: Percentage of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) returned from Pfam domain annotation of ALL (All individual sequences), CLS (Optimal clustering parameters), ASS (Assembled) and CLA (Clustered assembly) of simulated metatranscriptomes.

	LD			MD			HD		
	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec	Acc
All sequences	0.16	0.97	0.4	0.16	0.98	0.41	0.17	0.95	0.42
Optimal clustering	0.17	0.96	0.37	0.17	0.96	0.39	0.18	0.92	0.4
Assembly	0.29	0.93	0.47	0.3	0.94	0.49	0.21	0.93	0.44
Clustered assembly	0.3	0.92	0.46	0.31	0.91	0.47	0.21	0.91	0.42

Table 4.4: Performance summary of 454 simulations. Sensitivity, specificity and accuracy of Pfam annotations for: all sequences; cluster representatives transferred to members; assembly (contigs & singletons); and clustered assemblies for low diversity (LD), medium diversity (MD) and high diversity (HD) samples.

Although we do not know the precise taxonomic or functional composition of the two real metatranscriptomes, it is still interesting to note the effect these different approaches had on the frequency of protein domains. For the GIL data set, which we defined as a medium diversity sample, clustering only produces a small increase in the normalised abundance of domains compared to annotating all reads individually (see figure 4.3). However, assembly and a clustered assembly produce nearly double the number of domains per hundred thousand bases; this is very similar behaviour to the medium diversity simulated metatranscriptome. The higher diversity OMZ sample results were closer to the high diversity simulation, the main difference being that for the real data set, clustering produced more domains per hundred thousand bases than assembly (figure 4.3).

4.5 Discussion

In this chapter we have described the assessment of sequence processing methods (clustering, assembly and clustered assembly) using simulated metatranscriptome sequence data. We created simulated data sets to represent three

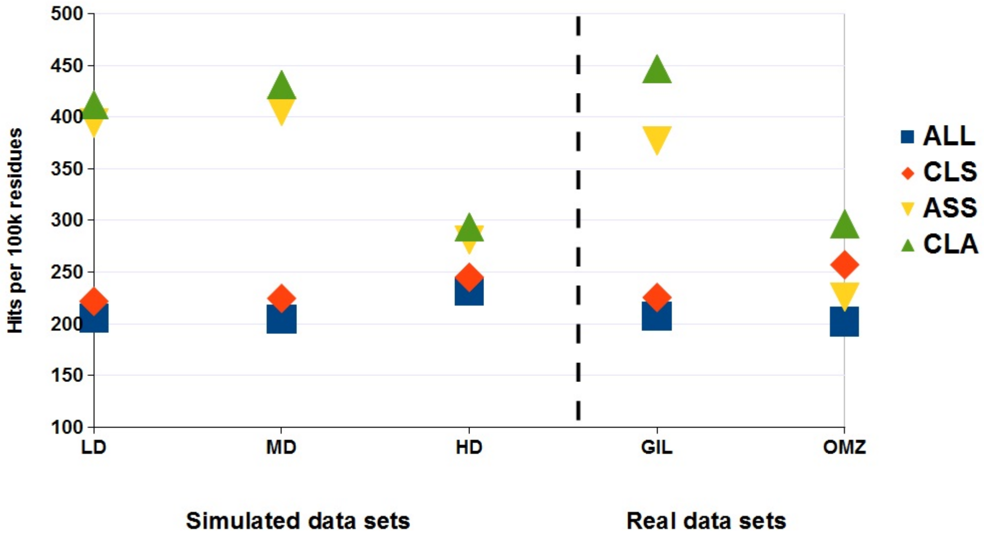


Figure 4.3: Normalised frequency of protein domains detected in simulated data sets (Low, (LD) Medium (MD) and High diversity (HD)) and real metatranscriptomes (GIL [Gilbert et al., 2008] and OMZ [Stewart et al., 2012]) for all individual annotated reads (ALL), optimal clustering parameters (CLS), assembled data (ASS) and clustered assembly (CLA).

different levels of taxonomic diversity using a power law distribution of transcript expression. Using popular clustering and assembly tools we compared the results of clustering on error free amino acid sequence fragments to those of translated error-prone 454 nucleotide sequences. Finally we compared the results of our simulations with those from real metatranscriptome data.

These results show that, of the sequence processing methods assessed, assembly produces the largest increase in protein domain detection and overall accuracy of the annotations. The longer reads produced allowing for more accurate classification of protein domains. Secondly, the results show that the level of diversity of the sample affects the quality of the assembly, in terms of the number of reads assembled, average contig length and in the resulting annotations. The greater the number of different taxa present in the sample, the lower the coverage of each organism and transcript will be, therefore reducing the likelihood of recreating full transcripts. The high diversity simulated data set assembled the worst and none of the processing methods assessed produced large increases in domain detection. For the real OMZ metatranscriptome, assembly actually proved to be inferior to clustering in terms of the number of domains detected.

The final choice as to whether to assemble or to cluster metatranscriptome sequences will ultimately depend on the diversity of the samples and the aims of any downstream analysis. It seems that little is to be gained from assembling highly diverse data sets, and rRNA sequencing or even microscopic observations could provide an estimate to species diversity.

Since this analysis was started, more metatranscriptome projects have employed Illumina sequencing platforms [Qi et al., 2011], [Mason et al., 2012], [Orsi et al., 2013]. This allows for a greater sequencing depth compared to 454 and although it requires more complicated data normalisation, it would be interesting to investigate the effect of different processing methods on this type of sequence data.

There are several limitations to the simulations we have performed. The composition of transcripts in a community of microbes is determined by a whole host of factors such as light, temperature, nutrient levels, season, sampling depth and interactions between species. In the approach we have employed, the number of transcript copies per genes is random, with each gene having an equal chance of being expressed at a certain level. In reality, transcripts from genes involved with fundamental processes - for example, biosynthesis or housekeeping genes - are more likely to be present at high or less variable levels.

Finally, note that we chose to use the clustering algorithm CD-HIT and the assembler MIRA (with uniform parameters for each sample) based on their suitability and popularity. However, other clustering applications such as BlastClust [Dondoshansky and Wolf, 2002] or UClust [Edgar, 2010] may provide better results. Comparing the results of the error free amino acid sequence annotations with the results of error containing translated nucleotide sequences highlight the need to either check all 6 reading frames or employ a more reliable reading frame detection method, such as FragGeneScan [Rho et al., 2010].

Chapter 5

Analysis of putative alleles in the polar diatom Fragilariopsis cylindrus

This chapter is adapted from ‘Adaptation to polar sea ice facilitated by allelic divergence in a psychrophilic eukaryote ’, T. Mock et al. in preparation.

5.1 Summary

In previous chapters we highlighted the pressing need for additional reference genomes for eukaryotic phytoplankton. One phytoplankton species, whose genome is currently in the draft stage is *Fragilariopsis cylindrus*, a keystone species in polar marine environments. This chapter describes the analysis of putative alleles (alternative forms of the same gene) in the draft genome sequence of *Fragilariopsis cylindrus*. We hypothesized that the cessation of sexual reproduction in *Fragilariopsis cylindrus* may have led to the high degree of heterozygosity observed, thereby allowing this species to adapt to extreme polar conditions. Using metatranscriptome sequences from *Fragilariopsis cylindrus*–like organisms and transcriptome sequences from two strains of a sexually reproducing eukaryotic phytoplankton, we investigated allelic variance in *Fragilariopsis cylindrus* genes based on their draft sequences to see if alleles from *Fragilariopsis cylindrus* were more divergent than homologous alleles in a sexually reproducing diatom. This project was a collaboration between the Schools of Computing Sciences and Environmental Sciences at UEA and the Stazione Zoologica Anton Dohrn in Napoli.

5.2 Background

Fragilariopsis cylindrus is a pennate diatom generally found in the polar seas and sea-ice. These psychrophilic (cold loving) organisms inhabit a harsh environment, which experiences large fluctuations in environmental conditions such as temperature, light and salinity and annual ice formation and melting change the environment drastically. Despite this, *F. cylindrus* thrives in such conditions and tends to dominate phytoplankton communities at the poles.

The genome of *F. cylindrus* was recently sequenced by the Joint Genome Institute (JGI) and the draft genome sequence along with annotation tracks for GO terms, KEGG pathways etc. is currently available (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html). The genome was sequenced using the Sanger platform at ∼7.25 fold coverage. The current draft of the genome consists of 271 scaffolds totalling ∼80 megabases. The genome contains around 5.4% gaps and at present it may be that some chromosomes are represented by multiple scaffolds and the precise genomic location of some scaffolds is unknown. The genome contains a total of 27,137 predicted genes. However, due to a high degree of heterozygosity in *F. cylindrus* [Strauss, 2012] (that is polymorphic regions of DNA at a specific locus) many regions of the genome could not be collapsed into a single consenus haplotype, and ∼30% of the predicted gene models are heterozygous gene copies present in more than one scaffold [Strauss, 2012].

Analysis of the heterozygous genes suggest that these are highly diverged alleles rather than gene duplications (paralogs) [Strauss, 2012]. Based on the assumption that paralogs would have a higher divergence than alleles, as paralogs would be free to evolve independently, the level of sequence similarity (>99% for the majority of pairs) between the heterozygous genes speaks for allelic pairs [Strauss, 2012]. The coverage of heterozygous gene copies was also investigated. Assuming that a heterozygous gene pair represents two copies of the same genomic locus, we might expect that each of the heterozygous genes has a twofold lower average coverage than paralogous genes. The results show that the primary gene copy (the copy on the larger of the two scaffolds, had a higher average coverage (∼6-fold) in comparison to the secondary copies ∼3-fold see figure 5.1. However, this is likely due to the method of assembly. As the primary gene copy is contained within the larger scaffold, it is likely that the coverage of the primary gene copy is higher as sequences that are identical or nearly identical to both would be recruited into the larger scaffold and it is only when the level of divergence exceeds a predefined threshold that the a new contig branches off from the first. Finally, analysis of scaffold synteny showed no evidence of large-scale genome duplication [Strauss, 2012].

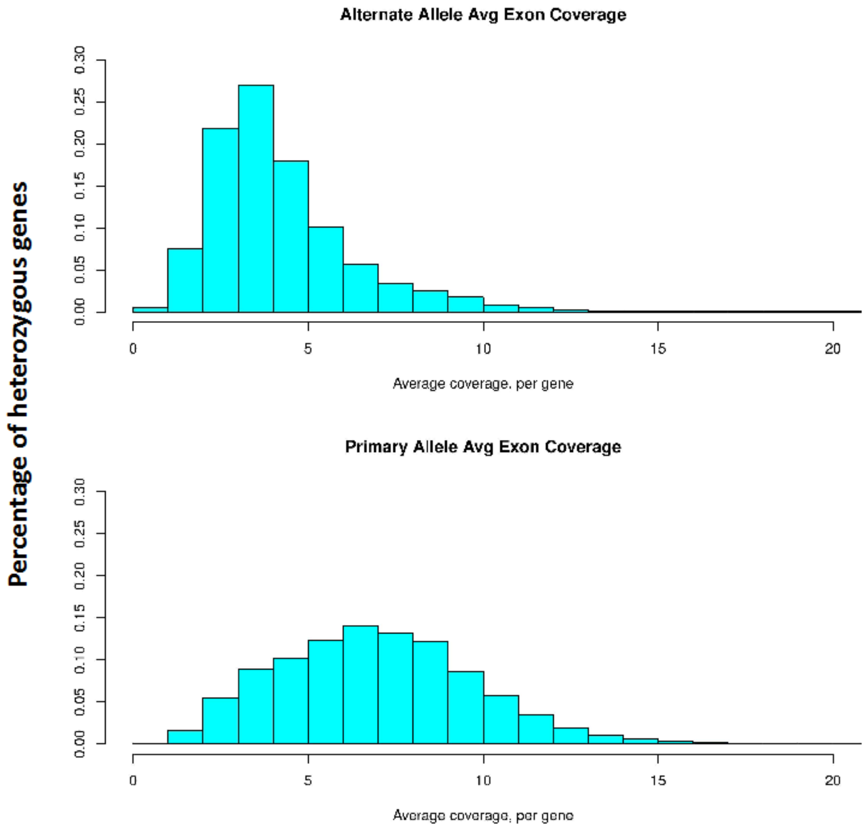


Figure 5.1: Average coverage of alternate (top) and primary (bottom) putative allele sequences of *Fragilariopsis cylindrus*. X-axis: average coverage per gene. Y:axis: percentage of total heterozygous gene sequences. Histogram supplied by JGI.

F. cylindrus is believed to have a clonal mode of reproduction as sexual reproduction has not been observed either in lab cultures or in the field, also several key meiotic genes are absent [Strauss, 2012]. Without Mendelian segregation and recombination of alleles during sexual reproduction, genomic heterozygosity and therefore allelic heterozygosity could increase with every generation. A transcriptome-wide RNA-SEQ analysis of *F. cylindrus* under different growth limitations revealed that more than half of the putative alleles showed unequal bi-allelic expression (>4 fold) [Strauss, 2012]. We hypothesized therefore that the lack of sexual reproduction in *F. cylindrus* could have been responsible for the high degree of heterozygosity, with the continuously diverging allelic pairs allowing *F. cylindrus* to adapt to extreme fluctuations in environmental conditions and inhabit such a niche environment. In order to test whether the putative *F. cylindrus* alleles are indeed more divergent we compared them with homologous genes from a mate pair (that is, the two strains were cross bred to produce the eventual strain to be sequenced for the genome sequence) the sexually reproducing diatom *Pseudonitzschia multistriata*.

5.3 Materials and methods

5.3.1 Culture preparation

The genome of *F. cylindrus* was derived from the CCMP1102 strain (`https://ncma.bigelow.org/ccmp1102`), which was grown at the University of Washington. Individual cells were isolated by flow cytometery and placed into separate wells of a culture plate (Thomas Mock, *personal communication*, 10/2/14).

The two strains of *Pseudonitzschia multistriata* were isolated from net samples collected in the Gulf of Naples. Individual cells were isolated under an inverted light microscope, cleaned and placed into separate wells of a culture plate. For more details see [Tesson et al., 2013].

5.3.2 Analysis of allelic transcripts from Fragilariopsis cylindrus–like organisms in a polar metatranscriptome

As our analysis of the Antarctic (ANT) metatranscriptome identified the majority of sequences as being most similar to *Fragilariopsis cylindrus* (see section 3.4.3), this provided us with an opportunity to investigate the expression of transcripts from allelic variants from a naturally occurring population. We extracted all bacillariophyta sequences (PhymmBL confidence score $\geq$0.9) whose best match was to *Fragilariopsis cylindrus* and aligned them to the full *Fragilariopsis cylindrus* amino acid gene set (including allelic variants). Alignments were performed using BLASTX (6 frame translation of nucleotide sequence query against amino acid reference), using soft masking, an e-value cut-off of $\leq$1e-10 and taking the best single hit per sequence.

5.3.3 Comparisons with a sexually reproducing diatom

Preparation of Fragilariopsis cylindrus transcripts

We first created our own transcript sequences for *Fragilariopsis cylindrus*, formed from only coding sequence regions (as we found that some JGI transcripts contained untranslated regions (UTRs)). We downloaded the latest FC scaffold sequences (Fracy1_assembly_scaffolds.fasta) and gff file (Fracy1_GeneModels_FilteredModels1.gff) from`http://genome.jgi-psf.org/Fracy1/Fracy1.download.ftp.html`. Then, based on a list of 9,062 allelic pairs supplied by JGI, we created a custom Perl script to extract, orientate and concatenate the nucleotide sequence for coding regions of each transcript from start codon to stop codon.

Preparation of Pseudonitzschia multistriata transcripts

Two strains (referred to as CIIO and CIIP from this point forwards) of *Pseudonitzschia multistriata*, a sexually reproducing pennate diatom were sequenced at the Joint Genome Institute using the Illumina platform. This produced approximately 92 million and 83 million, 150 bp paired end, strand specific reads for CIIO and CIIP, respectively. The data were filtered by quality score and adapter retention. Read pairs where at least one read matched the adapter sequences or showed a quality score of less than 30 for more than the 20% of the read were removed. This left approximately 25 million and 12 million quality filtered reads for CIIO and CIIP, respectively (Mariella Ferrante, Remo Sanges, *personal communication*, 10/11/12).

We assembled the quality filtered reads for CIIO and CIIP with Trinity [Grabherr et al., 2011] using a predetermined parameter set (Remo Sanges, *personal communication*, 13/11/12). To detect open reading frames (ORFs) in the assembled transcripts we used Transdecoder, (`http://transdecoder.sourceforge.net/`) a hexamer frequency, ORF detection utility contained in the Trinity download. Default parameters were used, except for lowering the minimum reading frame length to 50.

To identify allelic pairs from the *Pseudonitzschia multistriata* transcripts, we performed a reciprocal BLAST between the detected reading frames of the two assembled transcriptomes (BLASTN, overall identity $\geq$90%, requiring $\geq$75% coverage of both sequences). When a contig produced multiple candidate reading frames, we filtered the BLAST results to leave only the longest reading frame from each original transcript and its allelic counterpart.

Identifying orthologous alleles

The next step was to identify *Pseudonitzschia multistriata* alleles that were homologous to those in *Fragilariopsis cylindrus*. We performed reciprocal BLASTs between the putative alleles of each *Pseudonitzschia multistriata* strain and *Fragilariopsis cylindrus* alleles. The BLAST alignments were performed on the theoretical six frame translation of the sequences using TBLASTX, requiring $\geq$30% overall identity and $\geq$50% coverage of the query sequence (thresholds used in [Allen et al., 2008] to detect homologous transcripts). The BLAST results were combined and filtered to produce a list of *Pseudonitzschia multistriata* allelic pairs that either match to a singleton (i.e. a gene that has been collapsed into a single haplotype) or to the same allelic pair in *Fragilariopsis cylindrus*.

For each species, we then aligned each allelic pair using ClustalW2 (larkin2007) and compared the aligned sequences, base by base in parallel to calculate the number of SNPs (Single Nucleotide Polymorphisms), distinct indel (insertion/deletion) events and the number of positions included in indels.

Calculating divergence of allelic pairs

The following steps were performed for both *Fragilariopsis cylindrus* and *Pseudonitzschia multistriata*. For each sequence pair, we translated the nucleotide transcript sequences into amino acids and aligned them using Clustalw2 [Larkin et al., 2007], we also removed stop codons from the end of sequences if necessary. We then mapped the amino acid alignments back over the nucleotide sequences to ensure the nucleotide sequence contained full codons and were ‘in frame’. We then realigned the adjusted nucleotide sequences and calculated k_a/k_s for each sequence pair using codeml in pairwise mode as part of the

PAML 4.6 package [Yang, 2007]. For *Fragilariopsis cylindrus* singletons, k$_a$, k$_s$ and therefore k$_a$/k$_s$ were assumed to be zero.

5.4 Results

5.4.1 Allelic transcripts from Antarctic metatranscriptome

Of the 63,758 Antarctic metatranscriptome sequences identified as most similar to *Fragilariopsis cylindrus*, 41,130 aligned to *Fragilariopsis cylindrus* gene models with the thresholds used. Of these, 30,104 sequences ($\sim$73%) had their best match to a sequence from an allelic pair. We detected matches to a total of 455 allelic pairs. A great many of these pairs had only a single match to a single allelic variant, or very few ($<$5) in total. However, of the 167 allelic pairs with $\geq$5 matches to either allele, 78 pairs had significantly higher (Fisher's exact test, p-value $\leq$0.001) matches to one allele. Figure 5.2 shows the 20 most abundant allelic pairs (in terms of total number of sequences matching to both allelic variants). For all but one allelic pairs the number of sequences is significantly different. However, as the distribution of alleles in the Antarctic metatranscriptome is unknown. These results could be due to one putative allele being more abundant in the sample.

5.4.2 Comparisons with a sexually reproducing diatom

The CIIO and CIIP strains of *Pseudonitzschia multistriata* assembled into 39,714 and 32,198 contigs, respectively. We then retained any contigs where transdecoder predicted one or more candidate reading frames. This gave us 28,080

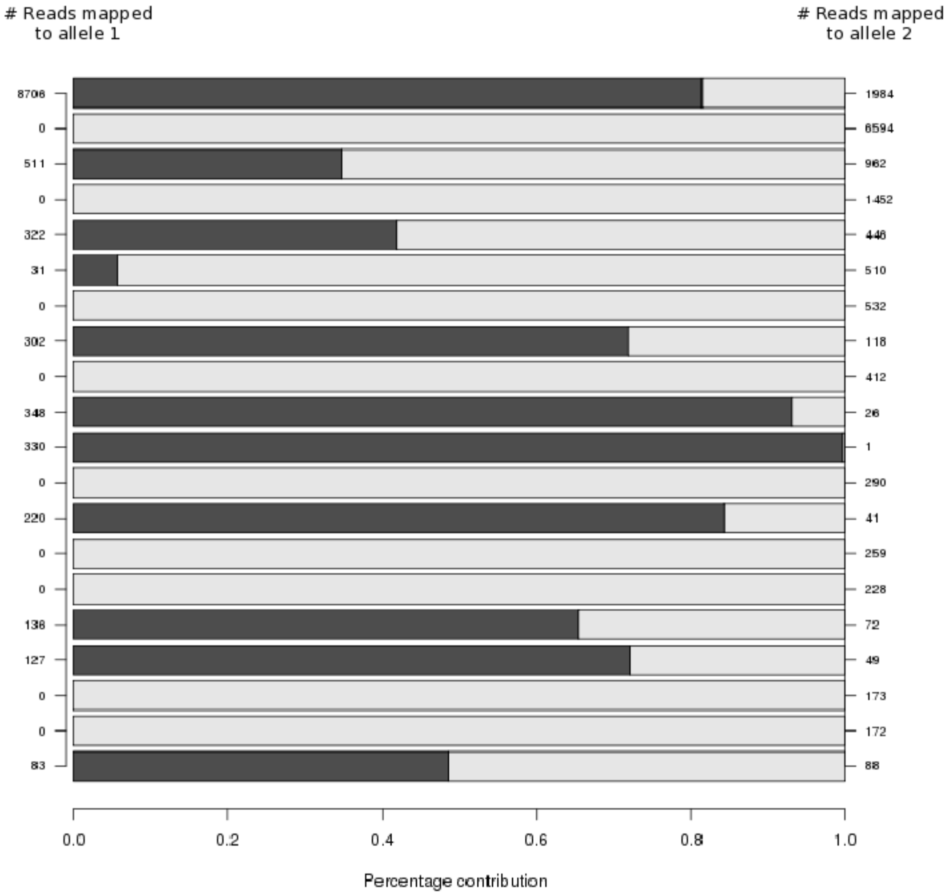


Figure 5.2: Percentage contribution of environmental transcripts to *Fragilariopsis cylindrus* allelic variants. Hits to allelic variant 1 (contained within the larger scaffold) are shown in dark grey along with the number of hits on the left Y-axis. Hits to allelic variant 2 (smaller scaffold) are shown in light grey with the number of hits shown on the right Y-axis. Diagram shows the twenty most abundant (total hit count to both allelic variants) allelic pairs.

and 24,486 contigs for CIIO and CIIP respectively. A reciprocal BLAST between the predicted reading frames produced 8,962 putative allelic pairs. The reciprocal BLAST between *Pseudonitzschia multistriata* and *Fragilariopsis cylindrus* identified 1,485 allelic pairs in *Pseudonitzschia multistriata* with homologs in *Fragilariopsis cylindrus* (with ∼64% matching to allelic pairs).

We decided to perform the reciprocal BLAST against all *Fragilariopsis cylin-*

drus transcripts rather than just the allelic pairs for two reasons. Firstly, by only using *Fragilariopsis cylindrus* alleles we could bias the analysis by only using the most heterozygous sequences, and secondly, this approach produced a greater number of sequence pairs to work with.

Despite the relatively stringent parameters used for allele detection in *Pseudonitzschia multistriata* we observed that a small number of allelic pairs aligned poorly during the k_a/k_s analysis (see figure 5.3). Upon closer inspection this was found to be due to the mapping of mistranslated reading frames over the original nucleotide sequences. The ClustalW2 alignment scores (overall % similarity) for the final adjusted nucleotide alignments were appended to the results file and allelic pairs aligning with <80% similarity were filtered out. This left a total of 1,354 homologous allelic pairs.

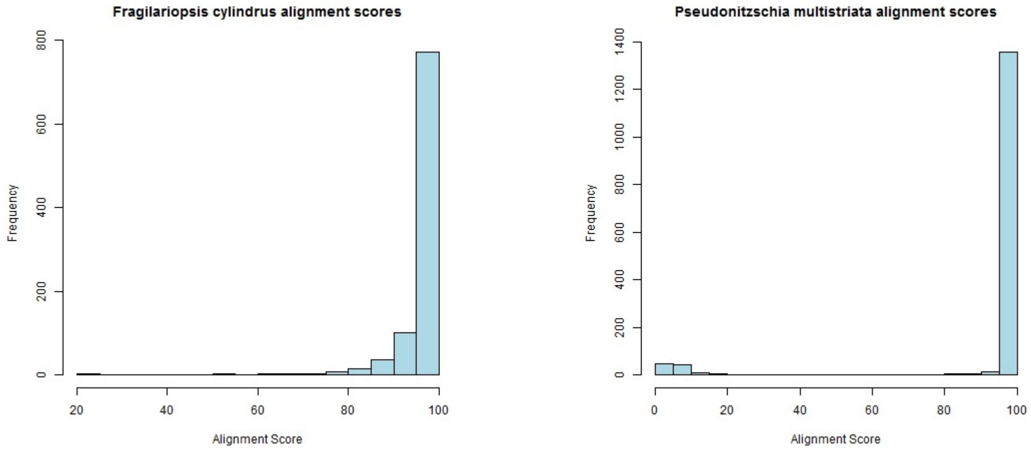


Figure 5.3: Histograms of Clustalw2 percentage alignment scores for allelic pairs for *Fragilariopsis cylindrus* (Left) and *Pseudonitzschia multistriata* (Right).

5.4.3 Interpretation of k_a/k_s data

From the sequence length and SNP count data produced in section 5.3.3, overall nucleotide polymorphism was calculated for each allelic pair of each species. Overall nucleotide polymorphism was higher, on average for *Fragilariopsis cylindrus* alleles than for homologous *Pseudonitzschia multistriata* alleles (see figure 5.4). The mean nucleotide polymorphism and standard error of mean in brackets were 0.013 (0.6310^{-3}) and 0.002 (0.1310^{-3}) for *Fragilariopsis cylindrus* and *Pseudonitzschia multistriata* respectively. This difference was deemed statistically significant by a paired T-test (p-value = 2.7810^{-55}).

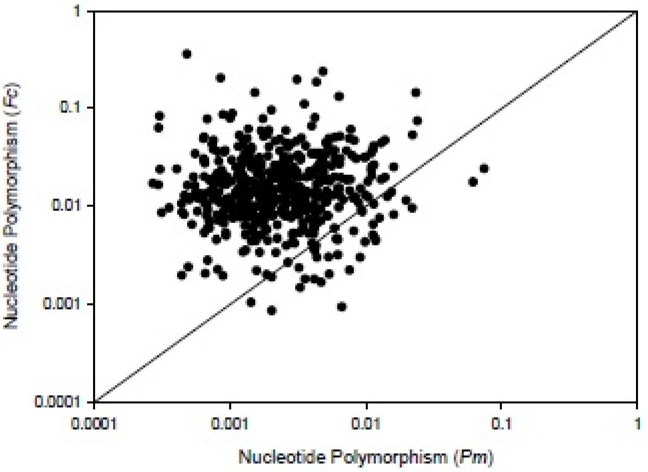


Figure 5.4: Scatter plot of nucleotide polymorphism between alleles of *Fragilariopsis cylindrus* (Fc) against nucleotide polymorphism of homologous putative *Pseudonitzschia multistriata* (Pm) alleles. (n = 1354).

However, contrary to our expectations, the average k_a/k_s ratio was lower for *Fragilariopsis cylindrus* (mean = 0.195; standard error of mean = 0.069), than for *Pseudonitzschia multistriata* (mean = 0.483; standard error of mean = 0.067). This could be for several reasons. The presence of singletons in the *Fragilariopsis cylindrus* data may have artificially lowered nucleotide divergence;

these sequences may have contained a low degree of divergence which has been lost as potential allelic variants were collapsed into single contigs. Secondly, partitioning nucleotide divergence into synonymous and non-synonymous mutations revealed that *Fragilariopsis cylindrus* alleles contained a greater number of both non-synonymous and synonymous mutations (see figure 5.5). So, although overall nucleotide divergence is indeed higher in *Fragilariopsis cylindrus*, the high number of synonymous mutations has likely eroded the signal of positive selection.

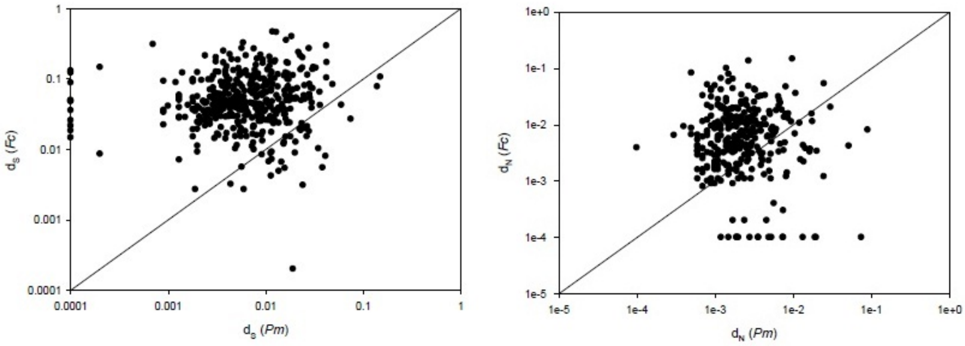


Figure 5.5: Plot of Synonymous nucleotide polymorphism (left), and non-synonymous nucleotide polymorphism (right) between alleles of *Fragilariopsis cylindrus* (Fc) and homologous putative *Pseudonitzschia multistriata* (Pm) alleles (n = 1354).

As synonymous mutations do not alter the encoded amino-acid sequence, they are not usually removed by purifying selection. Therefore, synonymous mutations are expected to accumulate almost linearly over time and the rate of synonymous mutations per synonymous site (k$_s$) can be used as a proxy for the age of a sequence and can, when plotted against the frequency of paralogous/allelic sequences be used to identify gene duplication events. According to [Lynch and Conery, 2000], duplicated genes begin with no polymorphism, but gradually acquire them over time. Thus, plotting the frequency of duplicated genes against their age (k_s) should produce an L-shaped plot with a large initial peak representing recently duplicated genes and decreasing exponentially over time (see [Blanc and Wolfe, 2004] figure 1A). Large-scale genome duplication events will greatly the increase the number of duplicated genes and lead to secondary peaks (see [Blanc and Wolfe, 2004] figure 1B), the number of synonymous mutations per synonymous site (k_s) these genes have subsequently acquired represents the age of the duplication event.

Figure 5.6 shows the expected L-shaped plot for *P. multistriata*, but, *F. cylindrus*, exhibits two peaks: the first of which most likely represents our singletons - where all polymorphism values are assumed to be zero; and a second which is indicative of a genome duplication event. However, as evidence from JGI suggests that these are allelic variants rather than paralogs (duplicates) this second peak may represent the point at which a large proportion of *F. cylindrus* alleles began diverging and may therefore indicate the point at which sexual reproduction ceased (Mark McMullan, *personal communication* 5/3/13).

5.5 Discussion

In this chapter we have described the analysis of highly diverged, putative allelic variants in the polar diatom *Fragilariopsis cylindrus*. To provide support for the hypothesis that the lack of sexual reproduction in *F. cylindrus* has facilitated the evolution of highly diverged alleles, allowing it to adapt to fluctuating environmental conditions we compared putative alleles of *F. cylindrus* with homologs in the sexually reproducing diatom *P. Multistriata* with the assumption that if *F. cylindrus* reproduces asexually it should exhibit a higher degree of divergence.

The results showed that overall nucleotide polymorphism was significantly

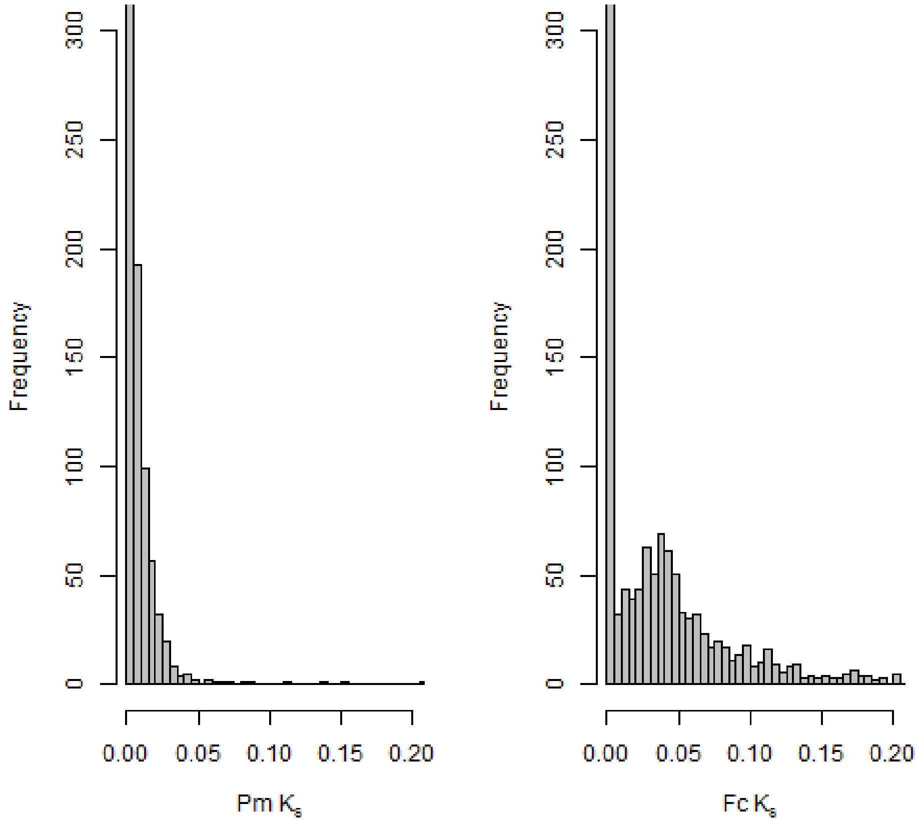


Figure 5.6: Left: Histogram of *Pseudonitzschia multistriata* putative allelic pair k$_s$ (synonymous substitutions per synonymous site) frequencies. Right: Histogram of *Fragilariopsis cylindrus* putative allelic pair k$_s$ frequencies

higher for *F. cylindrus*; that both synonymous and non-synonymous mutations were higher for *F. cylindrus*, even though the overall ratio of non-synonymous to synonymous mutations was lower. These results show that putative alleles in *F. cylindrus* are more divergent than their homologs in the sexually reproducing diatom *P. Multistriata*. Although this supports the hypothesis that cessation of sex in *F. cylindrus* led to a high degree of allelic variance, thereby allowing it to adapt to a fluctuating environment, there are other possible contributing factors for this high degree of heterozygosity to be considered: genome duplication and a large effective population size.

Analysis of the *F. cylindrus* genome by JGI suggest that it is unlikely that any large-scale genome duplication has occurred [Strauss, 2012], also, duplication events are not thought to be a major driver in the generation of diatom diversity [Bowler et al., 2008]. Another possibility is that a large effective population size has contributed towards this high degree of allelic variance, as in *Ciona intestinalis* [Dehal et al., 2002]. Work is ongoing by Cock Van Oosterhout producing simulations to assess the effect of effective population size on allelic diversity. A plot of the frequency of allelic variant pairs against synonymous mutations show a second peak in *F. cylindrus* which may represent a genome duplication event - which is not supported by the evidence from JGI - or this may coincide with the cessation of sexual reproduction.

Chapter 6

Discussion and future work

6.1 Summary

In Chapter 3 we described the computational pipeline that we set up to process metatranscriptome data from communities of eukaryotic phytoplankton from five representative marine environments. This involved a series of quality filtering and redundancy removal steps; identifying the taxonomic affiliations of transcript sequences and predicting their function; and finally a series of comparative and statistical analyses to identify similarities and differences between samples and investigate the relationship between environmental factors and the abundance of transcripts encoding for particular proteins.

The analysis revealed that all of the samples were dominated by eukaryotic phytoplankton and that the majority of transcripts encoded for proteins involved in biosynthesis. It also identified a correlation between the *in-situ* temperature of the sampling sites and the abundance of transcripts encoding for ribosomal proteins, a key component in the biosynthesis of new proteins. This was confirmed through laboratory experiments on model diatom species under control conditions. These experiments showed that lower temperatures lead to not only an increase in the level of transcription of ribosomal genes, but also an increase in the translation of ribosomal proteins. We theorised (and modelling simulations by a group in Exeter predicted) that this could alter the chemical composition of phytoplankton biomass (in terms of nitrate and phosphate) and have implications for fundamental marine biogeochemical cycles.

In Chapter 4 we described an assessment of sequence processing methods on simulated metatranscriptome data. We generated simulated 454 sequence data for microbial metatranscriptomes representing three different levels of taxonomic diversity. We then assessed the sensitivity, specificity and accuracy of protein domain annotation on the simulated sequence data sets using different processing methods: clustering; assembly; clustered assembly; and with no processing. The results showed that sequence assembly produced the largest increase in the overall accuracy of protein domain annotation, but that the benefits of assembly are reduced with higher levels of taxonomic diversity.

In Chapter 5 we described an analysis of the allelic diversity in the polar diatom *Fragilariopsis cylindrus*. To test the hypothesis that the cessation of sex in *Fragilariopsis cylindrus* led to a high degree of heterozygosity, thereby allowing it to adapt to an extreme environment, we performed comparative analyses with the sexually reproducing diatom *Pseudonitzschia multistriata*. Using transcriptome sequences from a mate pair of *Pseudonitzschia multistriata* strains, we investigated the rate of nucleotide divergence of *Fragilariopsis cylindrus* alleles compared to homologous allelic pairs of *Pseudonitzschia multistriata*. The results showed that overall nucleotide divergence and the number of both synonymous and nonsynonymous mutations were higher for *Fragilariopsis cylindrus* alleles than their *Pseudonitzschia multistriata* homologs, supporting our hypothesis.

6.2 Future work

6.2.1 Improvements to the metatranscriptome analysis

The major limitations of the analysis in Chapter 3 were the small number of sites sampled and the inconsistency in sampling methodologies. The samples were taken from disparate environments during different sampling cruises from different institutions. Therefore slight differences exist in the methods employed for water sampling, cell size filtering, mRNA sample preparation, the sequencing technology employed and inconsistencies in the recording of sampling site meta-data. These factors could lead to biased results and a low level of confidence in statistical testing. When testing relationships between environmental conditions and transcriptional activity, it is possible that other key differences in environmental conditions (for example the extremely high salinity of Antarctic sea-ice, or the high nutrient concentrations of Puget Sound) could skew statistical analyses.

In order to provide a more thorough investigation of the influence of a particular environmental factor (temperature) on the metabolism of eukaryotic phytoplankton with a greater degree of confidence, our intention is to perform a new metatranscriptome analysis. The aim is to produce over 100 metatranscriptome samples, using the latest Illumina RNA-SEQ sequencing technology, from phytoplankton communities across a latitudinal transect. This approach should allow us to analyse phytoplankton metabolism across a gradually changing temperature gradient with a high degree of confidence.

6.2.2 Illumina simulations for sequence processing assessment

As previously mentioned, the majority of early metatranscriptome projects employed 454 pyrosequencing. Recently however, there has been a shift towards the shorter sequence length, but significantly higher throughput of Illumina sequencing. Illumina sequencing can now produce reads exceeding 100bp and can provide around 15 times as much sequence data as the 454 price equivalent. The short reads provided by Illumina sequencing would require assembly and it would be of interest to extend the simulation approach described in Chapter 4 to produce simulated Illumina data to assess the current range of assembly programs designed for the problem of assembling transcripts from a mixed population of organisms.

6.3 Conclusions

Recent advances in high-throughput sequencing have allowed scientists unparalleled access to the genetic material of the previously unculturable majority of microbes. The *in-situ* sampling, sequencing and computational analysis of microbial metagenomes and metatranscriptomes has been applied to microbial communities from a diverse range of environments. These relatively new disciplines have quickly become important tools in, for example, the discovery of novel biocatalysts and in providing comparisons of microbial community metabolism and taxonomic make up.

Metatranscriptomics is necessarily a collaborative approach. The bioinformatics pipeline identified potentially interesting patterns in the data that molecular biologists were able to reproduce in specific species under laboratory conditions and investigate in more detail. Finally cell modelling techniques were used to make global predictions about the impact of this relationship on biogeochemical cycles.

The work described in this thesis highlights several current limitations to this field. Firstly, there is a need for more reference organisms in sequence databases. As previously stated, in most metatranscriptome projects, less than half of the sequence data return matches to known sequences. This may be partly due to sequence quality issues, the presence of untranslated regions, or possible non-coding RNAs, however, the need for more reference genomes and transcriptomes is clearly a limitation. At present we are limited to a relatively small number of model organisms, especially for the larger eukaryotic genomes.

While the majority of metatranscriptome analyses involve a series of similar processes, the set of tools employed, and parameters chosen vary greatly from project to project. For each step of a metatranscriptome analysis, including quality control, clustering, assembly, taxonomic classification, transcript function prediction and statistical tests, a large array of tools are available and there is no consensus as to how to analyse metatranscriptome data. The choice of tools and analyses employed will depend on the nature of the data, the type and diversity of organisms sequenced, which sequencing platform was employed, the amount of sequence data produced. This choice is also influence by the overall aims of the analysis (e.g. whether it is hypothesis driven or a more exploratory analysis), and the computational resources available. This type of analysis therefore requires a certain amount of flexibility and trial and error.

Despite the range of computational tools available for metatranscriptome analysis there remain many possible areas of improvement. The assembly of sequence fragments from a diverse range of organisms, each with varying coverage is still challenging. The taxonomic classification of sequence fragments is currently limited by sequence length and by the paucity of reference genomes. Also, the ever increasing scale of data that can be produced will necessitate improvements in data storage, algorithm speed and in summarising and visualising the results of large-scale metatranscriptome analysis.

The metatranscriptome analysis performed in this thesis also highlights the importance of experimental design. The sequence data represents the metabolic activity of a community of microbial organisms at the time of sampling. Slight variations in the methodology employed to obtain the samples could lead to slight biases in the eventual results. Great care should be taken to ensure that the methods employed for sampling, sequencing and analysing the data are as consistent as possible. The required meta-data should be carefully planned and recorded. Decisions should be made as to whether it is feasible to supplement the metatranscriptome data with metagenome data, 16s or 18s data or, producing transcriptomic or genomic data for key species in the environment of interest.

This is an exciting period for 'omics analyses. Each new analysis expands our understanding of microbe-environment interactions. Microbes from niche environments may reveal novel enzymes with potential medical, agricultural or industrial applications. As 'omics analyses become more and more a standard part of the microbiologists tool kit, bioinformatics faces the challenge of keeping up with the deluge of data and adapting to the goals of individual analyses.

Bibliography

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., and Roberts, K. (2002). *Molecular Biology of the Cell 4th edition*. National Center for Biotechnology InformationÕs Bookshelf. [cited at p. 12]

[Allen et al., 2008] Allen, A. E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., Finazzi, G., Fernie, A. R., and Bowler, C. (2008). Whole-cell response of the pennate diatom phaeodactylum tricornutum to iron starvation. *Proceedings of the National Academy of Sciences*, 105(30):10438–10443. [cited at p. 93]

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402. [cited at p. 24, 44]

[Amann et al., 1995] Amann, R. I., Ludwig, W., and Schleifer, K.-H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169. [cited at p. 5]

[Armbrust, 2009] Armbrust, E. (2009). The life of diatoms in the world's oceans. *Nature*, 459(7244):185–192. [cited at p. 10, 11, 55]

[Armbrust et al., 2004] Armbrust, E., Berges, J., Bowler, C., Green, B., Martinez, D., Putnam, N., Zhou, S., Allen, A., Apt, K., Bechner, M., et al. (2004). The genome of the diatom thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, 306(5693):79. [cited at p. 11]

[Ashburner et al., 2000] 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., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25. [cited at p. 34, 47]

[Bailly et al., 2007] Bailly, J., Fraissinet-Tachet, L., Verner, M.-C., Debaud, J.-C., Lemaire, M., Wésolowski-Louvel, M., and Marmeisse, R. (2007). Soil eukaryotic functional diversity, a metatranscriptomic approach. *The ISME journal*, 1(7):632–642. [cited at p. 6]

[Bairoch et al., 2005] Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl 1):D154–D159. [cited at p. 30, 32, 33]

[Baldrian et al., 2011] Baldrian, P., Kolařík, M., Štursová, M., Kopeckỳ, J., Valášková, V., Větrovskỳ, T., et al. (2011). Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME Journal*, 6(2):248–258. [cited at p. 6]

[Benson, 2011] Benson, D. (2011). l. karsch-mizrachi, dj lipman, j. ostell and ew sayers,genbank. *Nucleic Acids Res*, 39:D32–37. [cited at p. 32, 45]

[Bernstein et al., 2002] Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002). Global analysis of mrna decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays. *Proceedings of the National Academy of Sciences*, 99(15):9697–9702. [cited at p. 8]

[Birney et al., 2004] Birney, E., Clamp, M., and Durbin, R. (2004). Genewise and genomewise. *Genome research*, 14(5):988–995. [cited at p. 26]

[Blanc and Wolfe, 2004] Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell Online*, 16(7):1667–1678. [cited at p. 99]

[Booijink et al., 2010] Booijink, C. C., Boekhorst, J., Zoetendal, E. G., Smidt, H., Kleerebezem, M., and de Vos, W. M. (2010). Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. *Applied and environmental microbiology*, 76(16):5533–5540. [cited at p. 6]

[Bowler et al., 2008] Bowler, C., Allen, A., Badger, J., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R., et al. (2008). The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–244. [cited at p. 11, 101]

[Boyce et al., 2010] Boyce, D., Lewis, M., and Worm, B. (2010). Global phytoplankton decline over the past century. *Nature*, 466(7306):591–596. [cited at p. 9, 40]

[Bozarth et al., 2009] Bozarth, A., Maier, U., and Zauner, S. (2009). Diatoms in biotechnology: modern tools and applications. *Applied microbiology and biotechnology*, 82(2):195–201. [cited at p. 9]

[Bradbury, 2004] Bradbury, J. (2004). Nature's Nanotechnologists: Unveiling the Secrets of Diatoms. *PLoS Biology*, 2:1512–1514. [cited at p. 10]

[Brady and Salzberg, 2009] Brady, A. and Salzberg, S. (2009). Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673–676. [cited at p. 30, 45]

[Brady and Salzberg, 2011] Brady, A. and Salzberg, S. (2011). Phymmbl expanded: confidence scores, custom databases, parallelization and more. *Nature methods*, 8(5):367–367. [cited at p. 45]

[Brockman et al., 2008] Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W., Russ, C., Lander, E., Nusbaum, C., and Jaffe, D. (2008). Quality scores and snp detection in sequencing-by-synthesis systems. *Genome Research*, 18(5):763–770. [cited at p. 19]

[Chevreux et al., 2004] Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., and Suhai, S. (2004). Using the miraest assembler for reliable and automated mrna transcript assembly and snp detection in sequenced ests. *Genome research*, 14(6):1147–1159. [cited at p. 22, 72]

[Church, 2006] Church, G. (2006). Genomes for all. *Scientific American*, 294(1):46–54. [cited at p. 16]

[CLARKE, 1993] CLARKE, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1):117–143. [cited at p. 37]

[Conesa et al., 2005] Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676. [cited at p. 34]

[Crick et al., 1970] Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. [cited at p. 11]

[Crick, 1958] Crick, F. H. (1958). On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138. [cited at p. 11]

[Damon et al., 2012] Damon, C., Lehembre, F., Oger-Desfeux, C., Luis, P., Ranger, J., Fraissinet-Tachet, L., and Marmeisse, R. (2012). Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PloS one*, 7(1):e28967. [cited at p. 6]

[Dehal et al., 2002] Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., et al. (2002). The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–2167. [cited at p. 101]

[Denman, 2008] Denman, K. (2008). Climate change, ocean processes and ocean iron fertilization. *Marine Ecology Progress Series*. [cited at p. 10]

[DeSantis et al., 2006] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072. [cited at p. 21]

[Dohm et al., 2008] Dohm, J., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105. [cited at p. 17]

[Dondoshansky and Wolf, 2002] Dondoshansky, I. and Wolf, Y. (2002). Blastclust (ncbi software development toolkit). *NCBI, Bethesda, Md.* [cited at p. 85]

[Eddy, 2004] Eddy, S. R. (2004). What is dynamic programming? *Nature biotechnology*, 22(7):909–910. [cited at p. 23]

[Eddy et al., 2009] Eddy, S. R. et al. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, volume 23, pages 205–211. [cited at p. 24, 26, 27]

[Edgar, 2010] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461. [cited at p. 23, 85]

[Eppley, 1972] Eppley, R. W. (1972). Temperature and phytoplankton growth in the sea. *Fish. Bull*, 70(4):1063–1085. [cited at p. 40]

[Falgueras et al., 2010] Falgueras, J., Lara, A. J., Fernández-Pozo, N., Cantón, F. R., Pérez-Trabado, G., and Claros, M. G. (2010). Seqtrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics*, 11(1):38. [cited at p. 21]

[Falkowski et al., 1998] Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, 281(5374):200–206. [cited at p. 40]

[Federhen, 2012] Federhen, S. (2012). The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143. [cited at p. 46]

[Ferguson et al., 1976] Ferguson, R. L., Collier, A., and Meeter, D. A. (1976). Growth response ofthalassiosira pseudonana hasle and heimdal clone 3h to illumination, temperature and nitrogen source. *Chesapeake science*, 17(3):148–158. [cited at p. 61]

[Fernández-Arrojo et al., 2010] Fernández-Arrojo, L., Guazzaroni, M.-E., López-Cortés, N., Beloqui, A., and Ferrer, M. (2010). Metagenomic era for biocatalyst identification. *Current opinion in biotechnology*, 21(6):725–733. [cited at p. 5]

[Field et al., 1998] Field, C., Behrenfeld, M., Randerson, J., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237. [cited at p. 9]

[Fiers et al., 1976] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., et al. (1976). Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507. [cited at p. 13]

[Fiers et al., 1978] Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van De Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G., and Ysebaert, M. (1978). Complete nucleotide sequence of sv40 dna. *Nature*, 273(5658):113–120. [cited at p. 13]

[Finn et al., 2010] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The pfam protein families database. *Nucleic acids research*, 38(suppl 1):D211–D222. [cited at p. 26, 33, 46, 71]

[Fleischmann et al., 1995] Fleischmann, R., Adams, M., White, O., Clayton, R., Tatusov, R., Mushegian, A., Bork, P., Brown, N., Hayes, W., White, O., et al. (1995). Whole-genome random sequencing and assembly of haemophilus. *Science*, 269(5223):496–512. [cited at p. 13]

[Frias-Lopez et al., 2008] Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*, 105(10):3805–3810. [cited at p. 6, 8, 67, 68]

[Gerlach and Stoye, 2011] Gerlach, W. and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with carma3. *Nucleic acids research*, 39(14):e91–e91. [cited at p. 29]

[Gifford et al., 2010] Gifford, S. M., Sharma, S., Rinta-Kanto, J. M., and Moran, M. A. (2010). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME journal*, 5(3):461–472. [cited at p. 6, 7, 8, 67]

[Gilbert et al., 2008] Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., and Joint, I. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, 3(8):e3042. [cited at p. xiii, xv, xvi, xviii, 6, 7, 8, 68, 74, 76, 83, 145, 146]

[Gomez-Alvarez et al., 2009] Gomez-Alvarez, V., Teal, T. K., and Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *The ISME journal*, 3(11):1314–1317. [cited at p. 20]

[Gordon et al., 2009] Gordon, R., Losic, D., Tiffany, M., Nagy, S., and Sterrenburg, F. (2009). The glass menagerie: diatoms for novel applications in nanotechnology. *Trends in biotechnology*, 27(2):116–127. [cited at p. 10]

[Grabherr et al., 2011] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652. [cited at p. 22, 92]

[Guiry and Guiry, 2011] Guiry, M. and Guiry, G. (2011). Algaebase. world-wide electronic publication, national university of ireland, galway. [cited at p. 46]

[Haft et al., 2003] Haft, D. H., Selengut, J. D., and White, O. (2003). The tigrfams database of protein families. *Nucleic acids research*, 31(1):371–373. [cited at p. 26]

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919. [cited at p. 24]

[Hewson et al., 2009] Hewson, I., Poretsky, R. S., Dyhrman, S. T., Zielinski, B., White, A. E., Tripp, H. J., Montoya, J. P., and Zehr, J. P. (2009). Microbial community gene expression within colonies of the diazotroph, trichodesmium, from the southwest pacific ocean. *The ISME journal*, 3(11):1286–1300. [cited at p. 21]

[Huson et al., 2007] Huson, D., Auch, A., Qi, J., and Schuster, S. (2007). Megan analysis of metagenomic data. *Genome research*, 17(3):377–386. [cited at p. 29, 36]

[Institute, a] Institute, J. G. Fragilariopsis cylindrus genome page. http://genome.jgi-psf.org/Fracy1/Fracy1.home.html. [cited at p. 11]

[Institute, b] Institute, J. G. Pseudo-nitzschia multiseries genome page. http://genome.jgi.doe.gov/Psemu1/Psemu1.home.html. [cited at p. 11]

[Iseli et al., 1999] Iseli, C., Jongeneel, C. V., Bucher, P., et al. (1999). Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in est sequences. In *Proc Int Conf Intell Syst Mol Biol*, volume 7, pages 138–148. [cited at p. 26]

[John et al., 2009] John, D. E., Zielinski, B. L., and Paul, J. H. (2009). Creation of a pilot metatranscriptome library from eukaryotic plankton of a eutrophic bay(tampa bay, florida). *Limnology and Oceanography: Methods*, 7:249–259. [cited at p. 7, 44]

[Kariin and Burge, 1995] Kariin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, 11(7):283–290. [cited at p. 30]

[Klug and Michael, 1997] Klug, W. S. and Michael, R. (1997). Cummings. concepts of genetics. [cited at p. 11, 12, 14]

[Koski and Golding, 2001] Koski, L. and Golding, G. (2001). The closest blast hit is often not the nearest neighbor. *Journal of Molecular Evolution*, 52(6):540–542. [cited at p. 26, 31]

[Krause et al., 2008] Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A., and Stoye, J. (2008). Phylogenetic classification of short environmental dna fragments. *Nucleic acids research*, 36(7):2230–2239. [cited at p. 26]

[Kunin et al., 2008] Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4):557–578. [cited at p. 22, 68]

[Kuzio et al., 2006] Kuzio, J., Tatusov, R., and Lipman, D. (2006). Dust. *Unpublished but briefly described in: Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. Journal of Computational Biology*, 13(5):1028–1040. [cited at p. 43]

[Lagesen et al., 2007] Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007). Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Research*, 35(9):3100–3108. [cited at p. 21, 26]

[Larkin et al., 2007] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948. [cited at p. 93]

[Leininger et al., 2006] Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G., Prosser, J., Schuster, S., and Schleper, C. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, 442(7104):806–809. [cited at p. 6]

[Lesniewski et al., 2012] Lesniewski, R. A., Jain, S., Anantharaman, K., Schloss, P. D., and Dick, G. J. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *The ISME Journal*. [cited at p. 6]

[Letunic et al., 2008] Letunic, I., Yamada, T., Kanehisa, M., and Bork, P. (2008). ipath: interactive exploration of biochemical pathways and networks. *Trends in biochemical sciences*, 33(3):101–103. [cited at p. 34]

[Li et al., 2010] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272. [cited at p. 22]

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659. [cited at p. 23, 42, 71]

[Li et al., 2012] Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., et al. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1):25–37. [cited at p. 22]

[Liu et al., 2011] Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC genomics*, 12(Suppl 2):S4. [cited at p. 28, 29]

[Liu et al., 2012] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012. [cited at p. 17, 19]

[Lodish et al., 2000] Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Molecular cell biology. *New York*. [cited at p. 21]

[Loladze and Elser, 2011] Loladze, I. and Elser, J. J. (2011). The origins of the redfield nitrogen-to-phosphorus ratio are in a homoeostatic protein-to-rrna ratio. *Ecology letters*, 14(3):244–250. [cited at p. 65]

[Lozupone and Knight, 2005] Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235. [cited at p. 28]

[Ludwig et al., 2004] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., et al. (2004). Arb: a software environment for sequence data. *Nucleic acids research*, 32(4):1363–1371. [cited at p. 28]

[Lynch and Conery, 2000] Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155. [cited at p. 98]

[Lysholm et al., 2011] Lysholm, F., Andersson, B., and Persson, B. (2011). An efficient simulator of 454 data using configurable statistical models. *BMC research notes*, 4(1):449. [cited at p. 70]

[Maidak et al., 2001] Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker Jr, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., and Tiedje, J. M. (2001). The rdp-ii (ribosomal database project). *Nucleic acids research*, 29(1):173–174. [cited at p. 21]

[Manly, 2005] Manly, B. F. (2005). *Multivariate statistical methods: a primer*. Chapman & Hall. [cited at p. 58]

[Marchetti et al., 2012] Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E., and Armbrust, E. V. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, 109(6):E317–E325. [cited at p. 7, 8, 42]

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380. [cited at p. ix, 17, 18, 22]

[Markowitz et al., 2006] Markowitz, V., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., et al. (2006). The integrated microbial genomes (img) system. *Nucleic Acids Research*, 34(suppl 1):D344–D348. [cited at p. 70]

[Mason et al., 2012] Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S., Dubinsky, E. A., Fortney, J. L., Han, J., Holman, H.-Y. N., Hultman, J., Lamendella, R., et al. (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to deepwater horizon oil spill. *The ISME Journal*. [cited at p. 7, 8, 84]

[Mavromatis et al., 2007] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods*, 4(6):495–500. [cited at p. 68]

[Maxam and Gilbert, 1977] Maxam, A. and Gilbert, W. (1977). A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564. [cited at p. 16]

[McCormick and Cairns, 1994] McCormick, P. and Cairns, J. (1994). Algae as indicators of environmental change. *Journal of Applied Phycology*, 6(5):509–526. [cited at p. 9]

[McDonald, 2009] McDonald, J. H. (2009). *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore. [cited at p. 35]

[McHardy et al., 2006] McHardy, A., Martín, H., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2006). Accurate phylogenetic classification of variable-length dna fragments. *Nature methods*, 4(1):63–72. [cited at p. 30]

[McHardy and Rigoutsos, 2007] McHardy, A. C. and Rigoutsos, I. (2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Current opinion in microbiology*, 10(5):499–503. [cited at p. 31]

[Mende et al., 2012] Mende, D., Waller, A., Sunagawa, S., Järvelin, A., Chan, M., Arumugam, M., Raes, J., and Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one*, 7(2):e31386. [cited at p. 68]

[Meyer et al., 2008] Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386. [cited at p. 21, 23, 30]

[Mitra et al., 2010] Mitra, S., Gilbert, J. A., Field, D., and Huson, D. H. (2010). Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *The ISME journal*, 4(10):1236–1242. [cited at p. 29]

[Mitra et al., 2011] Mitra, S., Rupek, P., Richter, D., Urich, T., Gilbert, J., Meyer, F., Wilke, A., and Huson, D. (2011). Functional analysis of metagenomes and metatranscriptomes using seed and kegg. *BMC bioinformatics*, 12(Suppl 1):S21. [cited at p. 29]

[Mock et al., 2008] Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., BonDurant, S. S., Richmond, K., Rodesch, M., et al. (2008). Whole-genome expression profiling of the marine diatom thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences*, 105(5):1579–1584. [cited at p. 50]

[Moore et al., 2008] Moore, S. K., Mantua, N. J., Kellogg, J. P., and Newton, J. A. (2008). Local and large-scale climate forcing of. puget sound oceanographic properties on seasonal to interdecadal timescales. *Limnology and Oceanography*, 53(5):1746. [cited at p. 61]

[Moran, 2010] Moran, M. A. (2010). Metatranscriptomics: eavesdropping on complex microbial communities. *Issues*. [cited at p. viii, 5, 6, 7, 8]

[Moriya et al., 2007] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35(suppl 2):W182–W185. [cited at p. 33, 46]

[Mou et al., 2008] Mou, X., Sun, S., Edwards, R., Hodson, R., and Moran, M. (2008). Bacterial carbon processing by generalist species in the coastal ocean. *Nature*, 451(7179):708–711. [cited at p. 74]

[Nacher and Akutsu, 2006] Nacher, J. and Akutsu, T. (2006). Sensitivity of the power-law exponent in gene expression distribution to mrna decay rate. *Physics Letters A*, 360(1):174–178. [cited at p. 69]

[Namiki et al., 2011] Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2011). Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 116–124. ACM. [cited at p. 22]

[Niu et al., 2010] Niu, B., Fu, L., Sun, S., and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC bioinformatics*, 11(1):187. [cited at p. 20, 67]

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34. [cited at p. 33, 46, 71]

[Ogura et al., 2011] Ogura, A., Lin, M., Shigenobu, Y., Fujiwara, A., Ikeo, K., and Nagai, S. (2011). Effective gene collection from the metatranscriptome of marine microorganisms. *BMC genomics*, 12(Suppl 3):S15. [cited at p. 68]

[Orsi et al., 2013] Orsi, W. D., Edgcomb, V. P., Christman, G. D., and Biddle, J. F. (2013). Gene expression in the deep biosphere. *Nature*. [cited at p. 84]

[Parks and Beiko, 2010] Parks, D. H. and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–721. [cited at p. 36]

[Parsons et al., 1983] Parsons, T., Takahashi, M., and Hargrave, B. (1983). Biological oceanographic processes. *PERGAMON PRESS, OXFORD(UK). 1983.* [cited at p. 10]

[Peng et al., 2011] Peng, Y., Leung, H. C., Yiu, S., and Chin, F. Y. (2011). Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–i101. [cited at p. 22]

[Piganeau, 2012] Piganeau, G. (2012). *Genomic Insights into the Biology of Algae*, volume 63. Academic Press. [cited at p. 31]

[Pignatelli and Moya, 2011] Pignatelli, M. and Moya, A. (2011). Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PloS one*, 6(5):e19984. [cited at p. 68]

[Poretsky et al., 2009] Poretsky, R. S., Hewson, I., Sun, S., Allen, A. E., Zehr, J. P., and Moran, M. A. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the north pacific subtropical gyre. *Environmental microbiology*, 11(6):1358–1375. [cited at p. 8, 67, 68]

[Poroyko et al., 2010] Poroyko, V., White, J. R., Wang, M., Donovan, S., Alverdy, J., Liu, D. C., and Morowitz, M. J. (2010). Gut microbial gene expression in mother-fed and formula-fed piglets. *PloS one*, 5(8):e12459. [cited at p. 6]

[Pruesse et al., 2007] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21):7188–7196. [cited at p. 21, 30, 44]

[Prüfer et al., 2008] Prüfer, K., Stenzel, U., Dannemann, M., Green, R. E., Lachmann, M., and Kelso, J. (2008). Patman: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531. [cited at p. 43]

[Pruitt et al., 2012] Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic acids research*, 40(D1):D130–D135. [cited at p. 32]

[Pruitt et al., 2007] 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 research*, 35(suppl 1):D61–D65. [cited at p. 30, 71]

[Qi et al., 2011] Qi, M., Wang, P., O'Toole, N., Barboza, P. S., Ungerfeld, E., Leigh, M. B., Selinger, L. B., Butler, G., Tsang, A., McAllister, T. A., et al. (2011). Snapshot of the eukaryotic gene expression in muskoxen rumena metatranscriptomic approach. *PloS one*, 6(5):e20521. [cited at p. 7, 8, 67, 84]

[Raes et al., 2011] Raes, J., Letunic, I., Yamada, T., Jensen, L. J., and Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular systems biology*, 7(1). [cited at p. 37]

[Rho et al., 2010] Rho, M., Tang, H., and Ye, Y. (2010). Fraggenescan: predicting genes in short and error-prone reads. *Nucleic acids research*, 38(20):e191–e191. [cited at p. 85]

[Rinta-Kanto et al., 2012] Rinta-Kanto, J. M., Sun, S., Sharma, S., Kiene, R. P., and Moran, M. A. (2012). Bacterial community transcription patterns during a marine phytoplankton bloom. *Environmental Microbiology*, 14(1):228–239. [cited at p. 68]

[Rodriguez-Brito et al., 2006] Rodriguez-Brito, B., Rohwer, F., and Edwards, R. A. (2006). An application of statistics to comparative metagenomics. *BMC bioinformatics*, 7(1):162. [cited at p. 36]

[Ronaghi, 2001] Ronaghi, M. (2001). Pyrosequencing sheds light on dna sequencing. *Genome research*, 11(1):3–11. [cited at p. 17]

[Rosenow et al., 2001] Rosenow, C., Saxena, R. M., Durst, M., and Gingeras, T. R. (2001). Prokaryotic rna preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic acids research*, 29(22):e112–e112. [cited at p. 21]

[Rusch et al., 2007] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., et al. (2007). The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS biology*, 5(3):e77. [cited at p. 5, 37]

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467. [cited at p. 16]

[Schloss et al., 2009] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541. [cited at p. 28]

[Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864. [cited at p. 21]

[Schnoes et al., 2009] Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, 5(12):e1000605. [cited at p. 32]

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145. [cited at p. 16, 70, 75]

[Sherman et al., 2009] Sherman, B. T., Lempicki, R. A., et al. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13. [cited at p. 36]

[Sims et al., 2006] Sims, P., Mann, D., and Medlin, L. (2006). Evolution of the diatoms: insights from fossil, biological and molecular data. *Journal Information*, 45(4). [cited at p. 10]

[Smith et al., 2012] Smith, S. R., Abbriano, R. M., and Hildebrand, M. (2012). Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Research*, 1(1):2–16. [cited at p. viii, 10]

[Sorhannus, 2007] Sorhannus, U. (2007). A nuclear-encoded small-subunit ribosomal rna timescale for diatom evolution. *Marine Micropaleontology*, 65(1):1–12. [cited at p. 10]

[Stark et al., 2010] Stark, M., Berger, S., Stamatakis, A., and von Mering, C. (2010). Mltreemap-accurate maximum likelihood placement of environmental dna sequences into taxonomic and functional reference phylogenies. *BMC genomics*, 11(1):461. [cited at p. 29]

[Stewart et al., 2011] Stewart, F. J., Dmytrenko, O., DeLong, E. F., and Cavanaugh, C. M. (2011). Metatranscriptomic analysis of sulfur oxidation genes in the endosymbiont of solemya velum. *Frontiers in microbiology*, 2. [cited at p. 68]

[Stewart et al., 2010] Stewart, F. J., Ottesen, E. A., and DeLong, E. F. (2010). Development and quantitative analyses of a universal rrna-subtraction protocol for microbial metatranscriptomics. *The ISME journal*, 4(7):896–907. [cited at p. 8]

[Stewart et al., 2012] Stewart, F. J., Ulloa, O., and DeLong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environmental microbiology*, 14(1):23–40. [cited at p. xiii, xv, xvi, xviii, 8, 74, 76, 83, 145, 147]

[Strauss, 2012] Strauss, J. (2012). *A genomic analysis using RNA-SEQ to investigate the adaptation of the psychrophilic diatom Fragilariopsis cylindrus to the polar environment*. PhD thesis, School of Environmental Sciences, University of East Anglia, Norwich. [cited at p. 88, 89, 90, 101]

[Supek et al., 2011] Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800. [cited at p. 34]

[Tatusov et al., 2003] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., et al. (2003). The cog database: an updated version includes eukaryotes. *BMC bioinformatics*, 4(1):41. [cited at p. 34, 71]

[Teeling et al., 2004a] Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glöckner, F. O. (2004a). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947. [cited at p. 30]

[Teeling et al., 2004b] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. (2004b). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163. [cited at p. 30]

[Ter Braak, 1986] Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179. [cited at p. 49]

[Tesson et al., 2013] Tesson, S. V., Legrand, C., van Oosterhout, C., Montresor, M., Kooistra, W. H., and Procaccini, G. (2013). Mendelian inheritance pattern and high mutation rates of microsatellite alleles in the diatom¡ i¿ pseudo-nitzschia multistriata¡/i¿. *Protist*, 164(1):89–100. [cited at p. 90]

[Thomas et al., 2012] Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3. [cited at p. 23, 68]

[Tyson et al., 2004] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43. [cited at p. 69]

[Ueda et al., 2004] Ueda, H., Hayashi, S., Matsuyama, S., Yomo, T., Hashimoto, S., Kay, S., Hogenesch, J., and Iino, M. (2004). Universality and flexibility in gene expression from bacteria to human. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3765–3769. [cited at p. 69]

[Urich et al., 2008] Urich, T., Lanzén, A., Qi, J., Huson, D. H., Schleper, C., and Schuster, S. C. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*, 3(6):e2527. [cited at p. 6, 7, 68]

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science Signalling*, 291(5507):1304. [cited at p. 13]

[Von Der Haar, 2008] Von Der Haar, T. (2008). A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC systems biology*, 2(1):87. [cited at p. 21]

[Warnecke and Hess, 2009] Warnecke, F. and Hess, M. (2009). A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *Journal of biotechnology*, 142(1):91–95. [cited at p. 7]

[White et al., 2009] White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology*, 5(4):e1000352. [cited at p. 36]

[Whitman et al., 1998] Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583. [cited at p. 5, 6]

[Wommack et al., 2008] Wommack, K. E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: read length matters. *Applied and environmental microbiology*, 74(5):1453–1463. [cited at p. 8, 31]

[Wu et al., 2008] Wu, M., Eisen, J. A., et al. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):R151. [cited at p. 29]

[Xiong et al., 2012] Xiong, X., Frank, D. N., Robertson, C. E., Hung, S. S., Markle, J., Canty, A. J., McCoy, K. D., Macpherson, A. J., Poussier, P., Danska, J. S., et al. (2012). Generation and analysis of a mouse intestinal metatranscriptome through illumina based rna-sequencing. *PloS one*, 7(4):e36009. [cited at p. 6]

[Yang, 2007] Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591. [cited at p. 94]

[Ye et al., 2006] Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., et al. (2006). Wego: a web tool for plotting go annotations. *Nucleic acids research*, 34(suppl 2):W293–W297. [cited at p. 34]

[Ye and Doak, 2009] Ye, Y. and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS computational biology*, 5(8):e1000465. [cited at p. 34, 47]

[Yooseph et al., 2007] Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., et al. (2007). The sorcerer ii global ocean sampling expedition: expanding the universe of protein families. *PLoS biology*, 5(3):e16. [cited at p. 5, 37]

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829. [cited at p. 22]

Appendices

Appendix A

Appendix

A.1 Metatranscriptome sampling

The equatorial Pacific samples were taken from two stations: EPAC1: 0°, 155°W; EPAC2: 0°, 140°W during a cruise to the equatorial Pacific Ocean from 15th to 2nd of October 2006 onboard the RV Kilo Moana. The North Pacific sample was taken from one station (NPAC: 47°55.19 N; 122°20'38 W) during a Puget Sound cruise on the 15th of August 2007 onboard the Sorcerer (Craig Venter Institute, US). Water for RNA samples was pumped from about 8m depth onboard with a hose and peristaltic pump. Cells were immediately filtered onto autoclaved Nucleopore filters (25mm) with a pore size of 2μm. Not more than 500ml were filtered at a time in order to keep the filtration time $\prec$5 minutes per filter. Filters were subsequently flash frozen in liquid nitrogen and stored in the laboratory at -80°C. Phytoplankton were collected and concentrated by net tows from about 10m depth to the surface, using 0.25m diameter nets with a mesh size of 10 μm (Research Nets Inc. Redmond, WA, USA).

Antarctic samples were taken on two stations (ANT1: 65°06.11 S, 57°23.55 W; ANT2: 60°07.11 S, 47°54.55W) during the WWOS (Winter Weddell Outflow Study) cruise in Austral summer 2006 with the German Icebreaker Polarstern. Samples on ANT1 were obtained by icecore drilling and collecting microorganism communities from the lowermost cm (ice-waterinterface) of the ice core. For RNA extraction ice samples were melted in or washed with prefiltered (0.2μm) sea water or brine and cells were subsequently filtered onto Isopore filters (Millipore) (25mm) with a pore size of 1.2μm. Filters were subsequently flash frozen and stored in liquid nitrogen. Samples on ANT2 were obtained by fishing ice floes and collecting microorganisms from the ice-water interface as done on ANT1.

For the North Atlantic and Arctic, phytoplankton community samples were taken in June 2009 on board the RV Jan Mayan. Water samples at the DCM were taken directly from the CTD rosette (12.5 L Niskin bottles) in waters characterised as Arctic (June 20 SW Spitsbergen at 76° 36N; 18° 11E, temperature -1°C at 35 m) and Atlantic influenced (June 16 at the Polar front south of Bear Island at 73° 55N; 18° 46E, temperature +2.1°C at 50m). Cells were collected by filtration on 5μm pore-etched polycarbonate filters, flashfrozen in liquid nitrogen, and stored in a cryoshipper for transport to the laboratory.

A.2 Metatranscriptome sequencing

Several samples per station and ecosystem (EPAC, NPAC, ANT) were filtered onto 2μ pore size filters and subsequently flash frozen in liquid nitrogen and stored at -80°C. RNA extraction was performed with the ToTally RNA extraction kit (Ambion) according to the manufacturers recommendation. Eukaryotic mRNA was extracted with the Oligotex mRNA purification kit (Qiagen). The same kit was used to do an additional purification with the purified mRNA from the first time to reduce the contamination by rRNA and bacterial mRNA. Due to a very limited amount of double purified eukaryotic mRNA, we pooled all samples from each ecosystem (EPAC, NPAC, ANT). cDNA synthesis on the pooled samples was conducted with the SuperSmart PCR cDNA kit (Clontech) according to manufacturers recommendations. Libraries for next generation sequencing were constructed according to protocols for Roche 454 GS-FLX and GS-Titanium sequencing. GS-FLX sequencing was done at the NERC sequencing facility in Liverpool (UK) and GS-Titanium sequencing was done at Roche 454 (US).

For the NATL and ARC samples, extraction of total RNA from replicate filters was performed following standard protocols and a commercial kit (RNAeasy, Qiagen). Synthesis of full-length double-stranded cDNA (ds-cDNA) was performed from 250ng of total RNA of each sample (SMARTer PCR cDNA Synthesis Kit; Clontech) according to the manufacturer's instructions, allowing synthesis of full-length transcripts while maintaining the gene representation of unamplified samples. Full-length single-stranded DNA templates were then amplified by long-distance PCR using the Advantage 2 PCR Kit (Clontech). Replicate PCR reactions were performed for each library in order to obtain the amount required for sequencing (3 5μg), and subsequently pooled and purified using the MiniElute PCR Purification kit (Qiagen). The cDNA libraries were quantified using NanoDrop (ThermoScientific), and the quality of final samples was verified using agarose gel electrophoresis. Libraries were sequenced by a commercial service provider (BioCant, Portugal) using 454 FLX Titanium chemistry.

A.3 Eukaryotic organisms used for PhymmBL taxonomic classification

Phylum	Class	Order	Family	Genus	Species	Strain
Dinoflagellata	Dinophyceae	Gonyaulacales	Gonyaulacaceae	Alexandrium	catenella	NA
Bacillariophyta	Coscinodiscophyceae	Chaetocerotales	Chaetocerotaceae	Chaetoceros	neogracile	NA
Glaucophyta	Glaucophyceae	Glaucocystales	Glaucocystaceae	Cyanophora	paradoxa	NA
Ochrophyta	Phaeophyceae	Ectocarpales	Ectocarpaceae	Ectocarpus	siliculosus	NA
Haptophyta	Prymnesiophyceae	Isochrysidales	Noelaerhabdaceae	Emiliania	huxleyi	NA
Cryptophyta	Cryptophyceae	Pyrenomonadales	Geminigeraceae	Guillardia	theta	NA
Dinoflagellata	Dinophyceae	Peridiniales	Heterocapsaceae	Heterocapsa	triquetra	NA
Dinoflagellata	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Karenia	brevis	NA
Dinoflagellata	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Karlodinium	micrum	NA
Haptophyta	Pavlovophyceae	Pavlovales	Pavlovaceae	Pavlova	lutheri NA	

Table A.1: List of eukaryotic EST (Expressed Sequence Tag) libraries and their taxonomic classifications used for PhymmBL reference database. Taxonomic classifications taken from the NCBI taxonomy and Algaebase.

Phylum	Class	Order	Family	Genus	Species	Strain
Bacillariophyta	Coscinodiscophyceae	Thalassiosirales	Thalassiosiraceae	Thalassiosira	pseudonana	CCMP1335
Bacillariophyta	Bacillariophyceae	Naviculales	Phaeodactylaceae	Phaeodactylum	tricornutum	CCAP1055/1
Bacillariophyta	Bacillariophyceae	Bacillariales	Bacillariaceae	Fragilariopsis	cylindrus	NA
Ciliophora	Oligohymenophorea	Peniculida	Parameciidae	Paramecium	tetraurelia	sd4-2
Ciliophora	Oligohymenophorea	Hymenostomatida	Tetrahymenidae	Tetrahymena	thermophila	SB210
Apicomplexa	Coccidia	Eucoccidiorida	Cryptosporidiidae	Cryptosporidium	parvum	IowaII
Apicomplexa	Aconoidasida	NA	Theileriidae	Theileria	annulata	Ankara
Apicomplexa	Aconoidasida	Haemosporida	NA	Plasmodium	yoelii	17XNL
NA	Lobosa	Amoebida	Entamoebidae	Entamoeba	histolytica	HM-1:IMSS
Mycetozoa	Dictyostelia	Dictyosteliida	NA	Dictyostelium	discoideum	AX4
NA	NA	Choanoflagellida	Codonosigidae	Monosiga	brevicollis	MX1
Microsporidia	NA	NA	Unikaryonidae	Encephalitozoon	cuniculi	GB-M1
Basidiomycota	Tremellomycetes	Tremellales	Tremellaceae	Cryptococcus	neoformans	JEC21
Ascomycota	Schizosaccharomycetes	Schizosaccharomycetales	Schizosaccharomycetaceae	Schizosaccharomyces	pombe	972h
Ascomycota	Pezizomycetes	Pezizales	Tuberaceae	Tuber	melanosporum	Mel28
Ascomycota	Dothideomycetes	Pleosporales	Phaeosphaeriaceae	Phaeosphaeria	nodorum	SN15
Ascomycota	Eurotiomycetes	Eurotiales	Trichocomaceae	Aspergillus	nidulans	FGSC_A4
Ascomycota	LeotiomyceteS	Helotiales	Sclerotiniaceae	Sclerotinia	sclerotiorum	1980_UF-70
Ascomycota	Sordariomycetes	Sordariales	Sordariaceae	Neurospora	crassa	OR74A
Ascomycota	Saccharomycetes	Saccharomycetales	Saccharomycetaceae	Saccharomyces	cerevisiae	S288c
Rhodophyta	Cyanidiophyceae	Cyanidiales	Cyanidiaceae	Cyanidioschyzon	merolae	10D
Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	lucimarinus	CCE9901
Chlorophyta	Chlorophyceae	Chlamydomonadales	Volvocaceae	Volvox	carteri	f_nagariensis
Chlorophyta	Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae	Chlamydomonas	reinhardtii	NA
Streptophyta	Liliopsida	Poales	Poaceae	Oryza	sativa	japonica
Streptophyta	NA	Brassicales	Brassicaceae	Arabidopsis	thaliana	NA
Nematoda	Chromadorea	Rhabditida	Rhabditidae	Caenorhabditis	elegans	NA
Arthropoda	Branchiopoda	Diplostraca	Daphniidae	Daphnia	pulex	NA
Arthropoda	Insecta	Diptera	Drosophilidae	Drosophila	melanogaster	NA
Echinodermata	Echinoidea	Echinoida	Strongylocentrotidae	Strongylocentrotus	purpuratus	NA
Chordata	Actinopterygii	Cypriniformes	Cyprinidae	Danio	rerio	NA
Chordata	Mammalia	Primates	Hominidae	Homo	sapiens	GRCh37
Bacillariophyta	Pelagophyceae	Pelagomonadales	Pelagomonadaceae	Aureococcus	anophagefferens	NA
Chlorophyta	Mamiellophyceae	Mamiellales	Mamiellaceae	Micromonas	pusilla	CCMP1545

Table A.2: List of eukaryotic genomes and their taxonomic classifications used for PhymmBL reference database. Taxonomic classifications taken from the NCBI taxonomy and Algaebase.

A.4 Sampling site environmental conditions

Location	Temperature	NO_3	$Si(OH)_4$	PO_4	Salinity (PSU)	PAR(W/m^2)	Chl a (ug/L)	Day Length
ANT1	-2	7.8	6.1	2.2	39.3^S	0.02	93	13.73
ANT2	-2	n.d.	n.d.	n.d.	n.d.	0.02	n.d.	12.01
ARC	-1.1	5^W	2.5^W	0.5^W	34.2	2.19	n.d.	24
EPAC1	27	4.72	2.42	0.5	35.3^W	279.62	0.26	11.97
EPAC2	27	4.4	1.88	0.5	35.3^W	60.02	0.29	11.97
NATL	2.1	5^W	2.5^W	0.5^W	34.9	0.32	n.d.	24
NPAC	12	12.47	30.02	1.71	30	324.37	6.98^I	14.1

Table A.3: Temperature, nutrients given in mol/L, salinity (Practical Salinity Units), photosynthetically active radiation (PAR), , Chlorophyll a (Chl a) given in μg/L and day length. n.d. = no data available. S = Taken from 'SPINDLER, Michael. NEOGLOBOQUADRINA PACHYDERMA FROM ANTARCTIC SEA ICE. Proc. NIPR Symp. Polar Biol. Vol. 9. (1996)'. W = Data derived from World Ocean Atlas (Annual mean surface measurements). I= in situ flourescence.

A.5 North Pacific sample phytoplankton cell counts

Dominant phytoplankton	Cells/L
Coscinodiscus walesii	5472±894
Chaetoceros spp. single cells	13105±1983
Chaetoceros spp. chains	6280±453
Thalassiosira spp.	91129±7998
Thalassiosira nitzschioides	$\prec$1000
Pennate diatoms	$\prec$100
Dinoflagellates	$\prec$100
Unidentified flagellates	$\prec$1000

Table A.4: Taxonomic composition of major eukaryotic phytoplankton species in NPAC (North East Pacific (Puget Sound)) on 15th of August 2006. N=3.

A.6 Multiple correlation plot

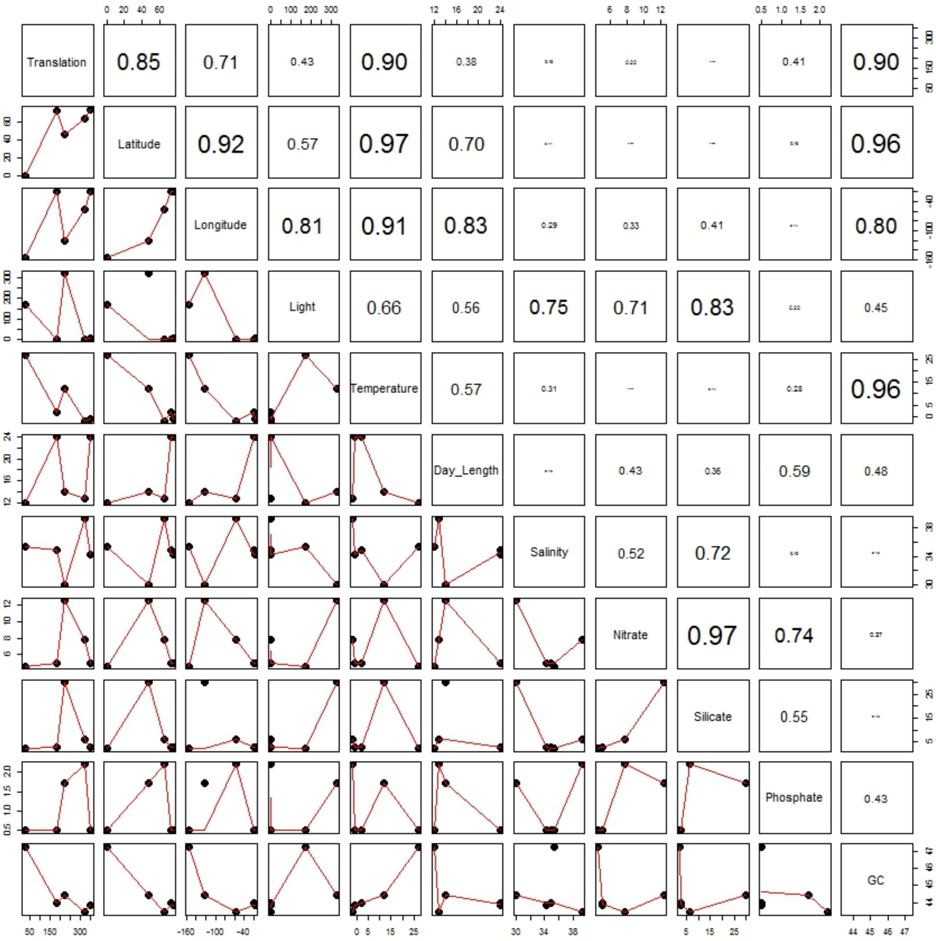


Figure A.1: Multiple correlation plot between normalised abundance of metatranscriptome sequences associated with GO:0006412 translation and environmental factors. Lower triangle displays scatter plot of factors from the central diagonal. Upper triangle displays scaled correlation coefficient between factors from central diagonal.

A.7 Additional CCA plots

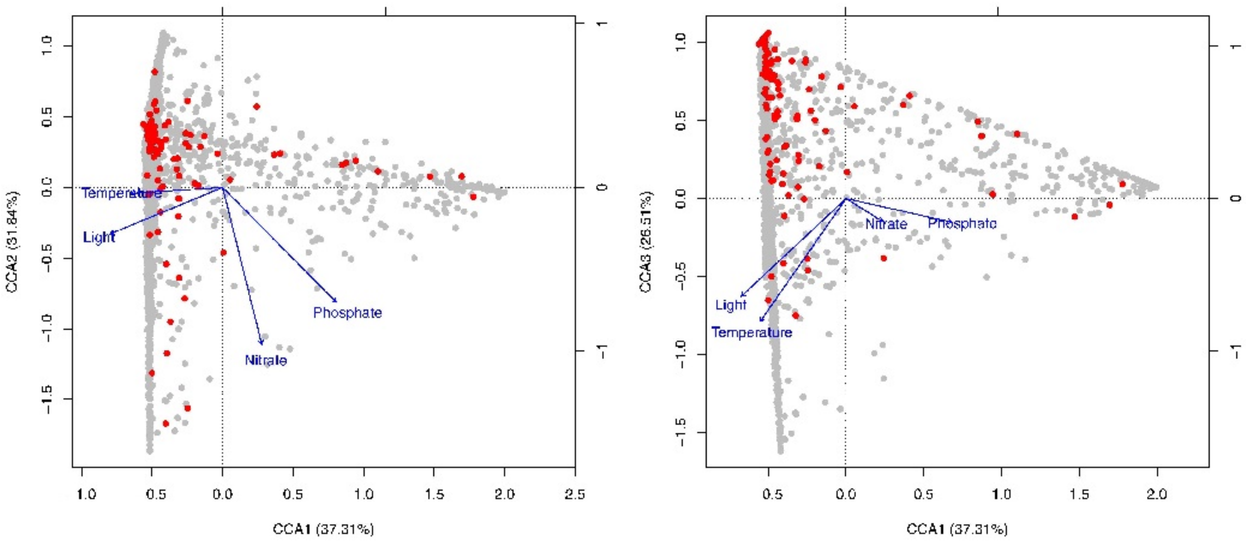


Figure A.2: Canonical correspondence analysis (CCA) between protein family (Pfam) abundance and environmental conditions (Temperature, Light, Nitrate and Phosphate) deduced from ocean samples in this study, red dots represent ribosomal transcripts. Figures represent dimensions 1 and 2 (Left) and 1 and 3 (Right).

A.8 Translational efficiency experiment

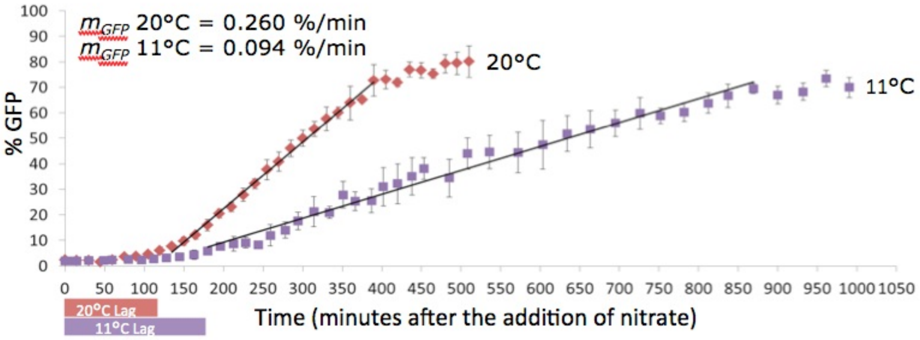


Figure A.3: Measurements of translation induction (lag time) and efficiency (slope) at different temperatures in *T. pseudonana* based on an inducible promoter (nitrate reductase) and measurements of % eGFP increase over time (N = 3; error bars denote standard deviation; mGFP = constant for % eGFP increase per minute).

Appendix B

Appendix B

B.1 Taxonomic composition of simulated metatranscriptomes

JGI-IMG ID	Organism	LD	MD	HD
640753001	Actinobacillus succinogenes 130Z	6	6	10
637000005	Alkalilimnicola ehrlichei MLHE-1	6	5	10
640753002	Alkaliphilus metalliredigens QYMF	5	4	9
646564504	Anabaena variabilis ATCC 29413	6	5	9
637000007	Anaeromyxobacter dehalogenans 2CP-C	6	5	10
639633006	Arthrobacter sp. FB24	6	5	10
643692004	Azotobacter vinelandii DJ, ATCC BAA-1303	6	5	9
643692007	Bacillus cereus 03BB102	7	5	11
645058795	Bifidobacterium longum DJO10A	5	4	9
640427103	Bradyrhizobium sp. BTAi1	56	111	10
643692011	Brevibacillus brevis NBRC 100599	5	4	8
641522608	Burkholderia ambifaria MC40-6	6	5	11
637000046	Burkholderia cenocepacia AU 1054	6	5	10
639633014	Burkholderia cenocepacia HI2424	6	5	10
637000051	Burkholderia sp. 383	7	5	11
640069307	Burkholderia vietnamiensis G4	6	5	10
637000053	Burkholderia xenovorans LB400	5	4	8
640427106	Caldicellulosiruptor saccharolyticus DSM 8903	7	5	10
637000160	Chelativorans sp. BNC1	5	4	9
637000072	Chlorobium chlorochromatii CaD3	5	6	10

JGI-IMG ID	Organism	LD	MD	HD
642555121	Chlorobium limicola DSM 245	6	5	9
639633020	Chlorobium phaeobacteroides DSM 266	5	5	9
641228485	Chloroflexus aurantiacus J-10-fl	7	5	10
637000075	Chromohalobacter salexigens DSM 3043	5	4	8
640753016	Clostridium beijerinckii NCIMB 8052	7	5	10
640069309	Clostridium thermocellum ATCC 27405	6	5	9
646311918	Cronobacter turicensis	7	4	11
637000087	Cytophaga hutchinsonii ATCC 33406	47	4	8
637000088	Dechloromonas aromatica RCB	5	4	9
641228488	Deinococcus geothermalis DSM 11300	6	5	10
643692021	Desulfobacterium autotrophicum HRM2, DSM 3382	6	4	9
637000095	Desulfovibrio desulfuricans G20	6	4	9
637000097	Ehrlichia canis Jake	7	6	9
637000098	Ehrlichia chaffeensis Arkansas	7	6	10
637000101	Enterococcus faecalis V583	5	4	7
641522626	Exiguobacterium sibiricum 255-15	6	5	10
640753026	Fervidobacterium nodosum Rt17-B1	5	4	8
637000116	Frankia sp. CcI3	6	5	10
641228492	Frankia sp. EAN1pec	6	5	10
637000119	Geobacter metallireducens GS-15	6	5	9
637000127	Histophilus somni 129PT	5	5	9
637000137	Jannaschia sp. CCS1	6	5	10
640753031	Kineococcus radiotolerans SRS30216	6	5	10
639633027	Lactobacillus brevis ATCC 367	4	4	8
639633028	Lactobacillus casei ATCC 334	6	5	9
639633029	Lactobacillus delbrueckii bulgaricus ATCC BAA-365	5	4	8
639633030	Lactobacillus gasseri ATCC 33323	6	5	11
640069315	Lactococcus lactis cremoris MG1363	5	5	8
639633034	Leuconostoc mesenteroides mesenteroides ATCC 8293	5	4	8
639633036	Magnetococcus sp. MC-1	5	4	10
639633037	Marinobacter aquaeolei VT8	6	5	10
637000161	Methanococcoides burtonii DSM 6242	5	5	10
637000162	Methanosarcina barkeri fusaro	5	4	9
637000164	Methanospirillum hungatei JF-1	6	5	10
637000165	Methylobacillus flagellatus KT	6	4	10
637000167	Moorella thermoacetica ATCC 39073	13	12	22
637000192	Nitrobacter hamburgensis X14	6	4	10
637000193	Nitrobacter winogradskyi Nb-255	6	4	10
637000194	Nitrosococcus oceani ATCC 19707	6	5	10
637000196	Nitrosomonas eutropha C71	6	5	10
637000197	Nitrosospira multiformis ATCC 25196	6	5	10
639633046	Nocardioides sp. JS614	6	5	10
640427126	Novosphingobium aromaticivorans DSM 12444	6	5	10
639633047	Oenococcus oeni PSU-1	4	4	8
639633048	Paracoccus denitrificans PD1222	5	5	10
639633049	Pediococcus pentosaceus ATCC 25745	5	4	9
644736398	Pedobacter heparinus DSM 2366	5	4	9

JGI-IMG ID	Organism	LD	MD	HD
637000204	Pelobacter carbinolicus DSM 2380	6	5	9
639633050	Pelobacter propionicus DSM 2379	6	5	10
642555146	Pelodictyon phaeoclathratiforme BU-1	6	5	10
637000208	Polaromonas sp. JS666	6	6	10
637000210	Prochlorococcus marinus MIT 9312	5	5	9
637000212	Prochlorococcus marinus NATL2A	7	4	10
642555149	Prosthecochloris aestuarii SK413, DSM 271	5	4	11
637000216	Pseudoalteromonas atlantica T6c	6	6	11
637000221	Pseudomonas fluorescens PfO-1	5	4	8
640427132	Pseudomonas putida F1	6	5	11
637000224	Pseudomonas syringae pv. syringae B728a	6	4	9
637000226	Psychrobacter arcticus 273-4	5	4	8
637000227	Psychrobacter cryohalolentis K5	7	5	10
643348570	Rhodobacter sphaeroides KD131	4	4	8
637000235	Rhodoferax ferrireducens T118	6	5	10
639279312	Rhodopseudomonas palustris BisA53	6	5	10
637000237	Rhodopseudomonas palustris BisB18	7	46	10
637000238	Rhodopseudomonas palustris BisB5	6	139	10
637000240	Rhodopseudomonas palustris HaA2	260	5	10
637000241	Rhodospirillum rubrum ATCC 11170	6	41	9
637000248	Rubrobacter xylanophilus DSM 9941	7	6	11
637000268	Ruegeria sp. TM1040	7	6	12
637000249	Saccharophagus degradans 2-40	5	5	10
639633057	Shewanella amazonensis SB2B	6	4	10
640069330	Shewanella baltica OS155	6	5	10
637000257	Shewanella frigidimarina NCIMB 400	6	5	10
640069331	Shewanella loihica PV-4	6	4	10
637000258	Shewanella oneidensis MR-1	6	5	10
639633058	Shewanella sp. ANA-3	7	5	10
637000260	Shewanella sp. MR-7	7	5	11
639633059	Shewanella sp. W3-18-1	5	5	10
637000271	Sphingopyxis alaskensis RB2256	6	5	10
644736409	Streptococcus suis SC84	6	5	9
639633062	Streptococcus thermophilus LMD-9	4	4	10
641522654	Synechococcus sp. PCC 7002	5	4	8
639633063	Syntrophobacter fumaroxidans MPOB	6	4	9
637000316	Syntrophomonas wolfei wolfei Goettingen	5	5	10
641522655	Thermoanaerobacter pseudethanolicus ATCC 33223	7	4	10
637000319	Thermobifida fusca YX	5	4	8
637000324	Thiobacillus denitrificans ATCC 25259	7	6	12
637000325	Thiomicrospira crunogena XCL-2	5	5	10
637000326	Thiomicrospira denitrificans ATCC 33889	6	6	8
637000329	Trichodesmium erythraeum IMS101	6	5	10
637000348	Xylella fastidiosa 9a5c	22	140	18
641522659	Xylella fastidiosa M12	10	68	4

Table B.1: List of organisms used for simulated metatranscriptomes, and copy numbers for Low Diversity (LD), Medium Diversity (MD) and High Diversity (HD) samples.

B.2 Assembly statistics

	LD	MD	HD	GIL	OMZ
#Contigs	24858	27752	26909	10903	42138
#Reads incorporated into contigs	115801 (46.32%)	125264 (50.11%)	74557 (29.82%)	73090 (47.62%)	168709 (53.88%)
Largest Contig (bp)	3011	3233	1368	11756	1175
Smallest Contig (bp)	42	41	40	50	40
Avg. Contig Length (bp)	298.6	298.3	257.3	415.5	244.4
#Debris reads	134199 (53.68%)	124736 (49.89%)	175443 (70.18%)	80397 (52.38%)	144438 (46.12%)

Table B.2: Summary of assembly statistics for Low Diversity (LD), medium diversity (MD), high diversity (HD) simulated metatranscriptomes, Gilbert et al.(GIL) [Gilbert et al., 2008] and Stewart et al. oxygen minimum zone (OMZ) [Stewart et al., 2012] real metatranscriptomes. All assemblies performed using MIRA in accurate, de novo, EST mode; non-uniform read depth, minimum of 2 reads per contig, and minimum contig length of 40.

B.3 Taxonomic composition of real metatranscriptomes

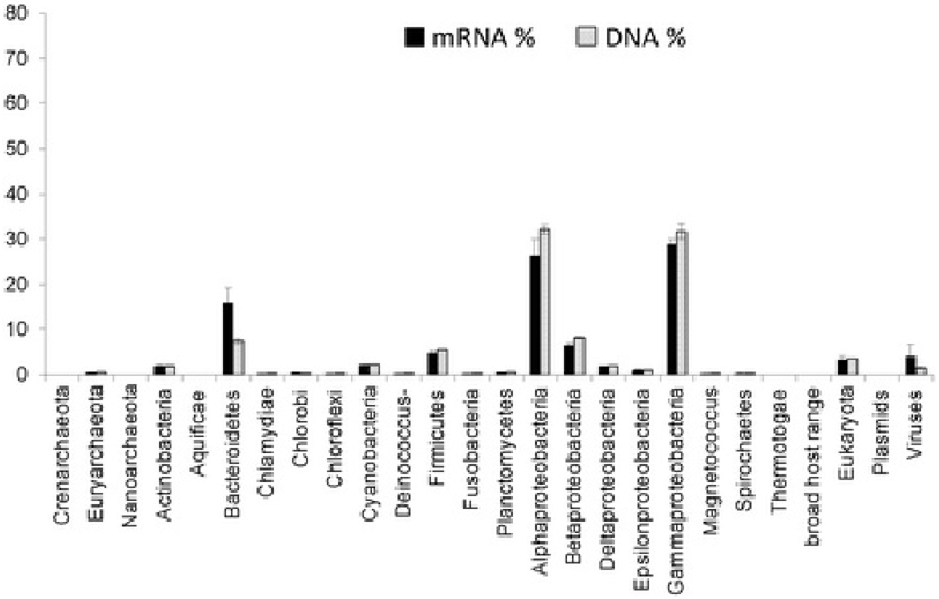


Figure B.1: Taxonomic affiliations of Gilbert mid-bloom metatranscriptome (Black) and metagenome (Grey) sequences. Figure reproduced from [Gilbert et al., 2008].

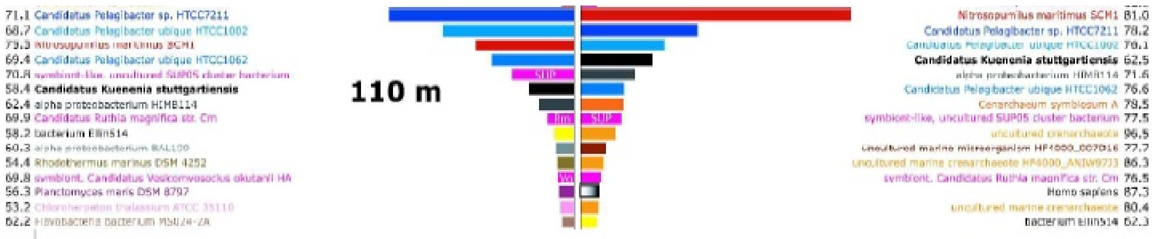


Figure B.2: Taxonomic affiliations of Stewart 110m metagenome (Left) and metatranscriptome (Right) sequences. Part of a figure reproduced from [Stewart et al., 2012].