Bioinformatic Methods in Metagenomics

Johannes Alneberg

Doctoral Thesis, 2018
KTH Royal Institute of Technology
Engineering Sciences in Chemistry,
Biotechnology and Health
Department of Gene Technology
Science for Life Laboratory
SE-171 65 Solna, Sweden

KTH Royal Institute of Technology
Engineering Sciences in Chemistry,
Biotechnology and Health
Department of Gene Technology

Science for Life Laboratory
SE-171 65 Solna
Sweden

Printed by Universitetsservice US-AB
Drottning Kristinas väg 53B
SE-100 44 Stockholm
Sweden

ISBN: 978-91-7729-799-4
TRITA-CBH-FOU-2018:25

Abstract

Microbial organisms are a vital part of our global ecosystem. Yet, our knowledge of them is still lacking. Direct sequencing of microbial communities, i.e. metagenomics, have enabled detailed studies of these microscopic organisms by inspection of their DNA sequences without the need to culture them. Furthermore, the development of modern high-throughput sequencing technologies have made this approach more powerful and cost-effective. Taken together, this has shifted the field of microbiology from previously being centered around microscopy and culturing studies, to largely consist of computational analyses of DNA sequences. One such computational analysis which is the main focus of this thesis, aims at reconstruction of the complete DNA sequence of an organism, i.e. its genome, directly from short metagenomic sequences.

This thesis consists of an introduction to the subject followed by five papers. Paper I describes a large metagenomic data resource spanning the Baltic Sea microbial communities. This dataset is complemented with a web-interface allowing researchers to easily extract and visualize detailed information. Paper II introduces a bioinformatic method which is able to reconstruct genomes from metagenomic data. This method, which is termed CONCOCT, is applied on Baltic Sea metagenomics data in Paper III and Paper V. This enabled the reconstruction of a large number of genomes. Analysis of these genomes in Paper III led to the proposal of, and evidence for, a global brackish microbiome. Paper IV presents a comparison between genomes reconstructed from metagenomes with single-cell sequenced genomes. This further validated the technique presented in Paper II as it was found to produce larger and more complete genomes than single-cell sequencing.

Keywords

Bioinformatics, Metagenomics, Microbiome, Binning, Baltic Sea

Sammanfattning

Mikrobiella organismer är en vital del av vårt globala ekosystem. Trots detta är vår kunskap om dessa fortfarande begränsad. Sekvensering direkt applicerad på mikrobiella samhällen, så kallad metagenomik, har möjliggjort detaljerade studier av dessa mikroskopiska organismer genom deras DNA-sekvenser. Utvecklingen av modern sekvenseringsteknik har vidare gjort denna strategi både mer kraftfull och mer kostnadseffektiv. Sammantaget har detta förändrat mikrobiologi-fältet, från att ha varit centrerat kring mikroskopi, till att till stor del bero på dataintensiva analyser av DNA-sekvenser. En sådan analys, som är det huvudsakliga fokuset för den här avhandlingen, syftar till att återskapa den kompletta DNA-sekvensen för en organism, dvs. dess genom, direkt från korta metagenom-sekvenser.

Den här avhandlingen består av en introduktion till ämnet, följt av fem artiklar. Artikel I beskriver en omfattande databas för metagenomik över Östersjöns mikrobiella samhällen. Till denna databas hör också en webbsida som ger forskare möjlighet att lätt extrahera och visualisera detaljerad information. Artikel II introducerar en bioinformatisk metod som kan återskapa genom från metagenom. Denna metod, som kallas CONCOCT, används för data från Östersjön i artikel III och Artikel V. Detta möjliggjorde återskapandet av ett stort antal genom. Analys av dessa genom presenterad i Artikel III ledde till hypotesen om, och belägg för, ett globalt brackvattenmikrobiom. Artikel IV innehåller en jämförelse mellan genom återskapade från metagenom och individuellt sekvenserade genom. Detta validerade metoden som presenterades i Artikel II ytterligare då denna metod visade sig producera större och mer kompletta genom än sekvensering av individuella celler.

Nyckelord

Bioinformatik, Metagenomik, Mikrobiom, Binning, Östersjön

List of publications

I. Johannes Alneberg, John Sundh, Christin Bennke, Sara Beier, Daniel Lundin, Luisa W. Hugerth, Jarone Pinhassi, Veljo Kisand, Lasse Riemann, Klaus Jürgens, Matthias Labrenz, Anders F. Andersson. *BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea*. Scientific Data (in press).

II. Johannes Alneberg*, Brynjar Smári Bjarnason*, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, Christopher Quince. (2014). *Binning metagenomic contigs by coverage and composition*. Nature Methods volume 11, pages 1144–1146 doi:10.1038/nmeth.3103

III. Luisa W. Hugerth, John Larsson, Johannes Alneberg, Markus V. Lindh, Catherine Legrand, Jarone Pinhassi, Anders F. Andersson. (2015). *Metagenome-assembled genomes uncover a global brackish microbiome*. Genome Biology 16:279 doi:10.1186/s13059-015-0834-7

IV. Johannes Alneberg*, Christofer M.G. Karlsson*, Anna-Maria Divne, Claudia Bergin, Felix Homa, Markus V. Lindh, Luisa W. Hugerth, Thijs JG Ettema, Stefan Bertilsson, Anders F. Andersson, Jarone Pinhassi. *Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes*. Manuscript in review

V. Johannes Alneberg, Christin Bennke, Sara Beier, Jarone Pinhassi, Klaus Jürgens, Martin Ekman, Karolina Ininbergs, Matthias Labrenz, Anders F. Andersson. *Recovering 2,032 Baltic Sea microbial genomes by optimized metagenomic binning*. Manuscript

*These authors contributed equally to respective paper.

Related publications

- Olov Svartström, Johannes Alneberg, Nicolas Terrapon, Vincent Lombard, Ino de Bruijn, Jonas Malmsten, Ann-Marie Dalin, Emilie EL Muller, Pranjul Shah, Paul Wilmes, Bernard Henrissat, Henrik Aspeborg, Anders F Andersson. (2017). *Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation*. The ISME Journal volume 11, pages 2538–2551 doi:10.1038/ismej.2017.108

- Christopher Quince, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins and A. Murat Eren. *DESMAN: a new tool for de novo extraction of strains from metagenomes*. (2017). Genome Biology 18:181 doi:10.1186/s13059-017-1309-9

Table of Contents

Introduction

On our planet, cellular life is present almost everywhere. Microbes thrive in environments as hostile as the acidic runoff water from a mine to the nutritious and protected environment of inside your gut. Furthermore, since all currently known cellular life forms are DNA based, environmental DNA is therefore present wherever you look for it.

While some parts of the DNA sequence in a cell have been reasonably conserved for several hundred million years, other parts of the same sequence might be unique to that individual cell due to novel mutations. This enables us to use DNA to characterize the microbes present in a certain environment: to find out who is there and what proportion of the community that they constitute. But DNA is far from only useful for this kind of fingerprinting, it also encodes the full capability inherent to the cell.

This thesis focuses on computational methods to process environmental DNA sequences. A special focus is on approaches to reconstruct the complete DNA sequence for species present in the community. Furthermore, most data studied will be from the Baltic Sea. Besides a short introduction, the main content of this thesis consists of a number of articles and manuscripts that I will refer to in this introduction as papers I, II, III, IV and V respectively. Paper I presents a processed dataset for the Baltic Sea together with a web based interface. Paper II presents a general method to reconstruct complete microbial DNA sequences using multiple environmental samples. Paper III uses this very method to investigate a Baltic Sea dataset. Paper IV compares two commonly used methods for DNA sequence reconstruction. Finally, paper V extends paper III with a new substantially larger dataset.

Primer

To a molecular biologist, it is perhaps truly offensive to say that DNA consists of the letters A, C, G, and T: ignoring the molecular structure and not spelling out the names of the nucleotides that these abbreviations represent (the names are Adenine, Thymine, Cytosine and Guanine. I don’t want to offend anyone). But within bioinformatics, the field to where this thesis belongs, this is a very useful abstraction in order to transcribe molecular information into a plain text file or into more complex data structures which can be easily accessible by a computer. Therefore, this will be the starting point of this thesis. While DNA is the main carrier of hereditary information for all cellular life, I will not try to elaborate on how this is accomplished. Furthermore, I will not attempt to explain the dynamics between DNA, RNA, proteins or any other of the important molecules of the cell. Instead, our starting point will be the following definitions which are chosen in order to reflect the common use within the field and not necessarily the most scientifically precise:

- DNA sequence: A sequence of any of the letters A, C, G and T.
- Gene: A DNA segment which is predicted to encode for a certain RNA or Protein. When encoding for a protein, the gene uses a messenger RNA (mRNA) as an intermediate stage.
- Genome: The complete DNA sequence of a cell.

Furthermore, even the most molecularly ignorant bioinformatician needs to know that each DNA sequence has exactly one complementary sequence where all occurrences of A:s are paired with T:s and all C:s are paired with G:s and vice versa. This complementary sequence is always given in the reverse order and is termed the reverse complement.

Since this thesis is dedicated to the post-processing of sequences produced by sequencing machines, some specific knowledge about these sequences are necessary. The output from an Illumina sequencing machine, which have been used for all papers included in this thesis, are millions of relatively short sequences called reads. The reads are normally paired, where the two member reads of the pair originate from different ends of the same molecule. This is called paired-end sequencing. Furthermore, the reads which have been used in papers included in this thesis were of length 100 base pairs (bp) for papers II & III while papers I,IV & V also includes some runs with 125 bp reads. Furthermore, paper IV contained reads of length 300 bp for a special application. All but the 300 bp reads were produced by the Illumina HiSeq machine.

Environmental microbiology

Most people outside of science likely associate the word bacteria or microbes with diseases. These disease causing microbes are so called pathogens. However, for the last 30 years or so, an ever increasing scientific attention have been directed towards commensal and symbiotic microbes (Marchesi 2011). These are the names of microbes which instead coexist in peace or collaborate with its host. The increase in attention can to a large extent be attributed to metabarcoding and metagenomics, two methods that will be presented later in this thesis.

With the exception of paper II, where human associated data will be used to showcase the method presented, this thesis will neither focus on pathogenic nor human associated microbes. Instead, the focus will be on microbes in the environment, the world's most diverse group of organisms, and more specifically microbes living in the Baltic Sea. By definition, most microbes are invisible to the naked eye. I will therefore start this section with an example about phytoplankton, the photosynthesizing microbes of oceans and lakes, to illustrate the major ecological contribution by microbes.

While it is easy to understand the importance of plants as major primary producers on land, the importance of the major primary producers of the oceans was underestimated for a long time. Given the size of individual microbes, it is rather contradictory that scientists had to use space satellites in order to reliably estimate their global importance (Falkowski 2012). The results were nevertheless stunning: microbial organisms of the ocean account for almost the same amount of carbon uptake and oxygen gas generation, as do plants (Falkowski 2012).

While phytoplanktons, that can be either single-cell eukaryotes or prokaryotes (i.e. cyanobacteria), are among the most important microbes in the ocean, they are far from the only ones. On the contrary, extrapolated measurements of cell counts for ocean water samples showed that prokaryotic phytoplankton only accounted for a few percent of the total number of prokaryotic cells in the oceans (Whitman, Coleman, and Wiebe 1998). From those estimates, it was also found that all prokaryotes together carried around 10 times the amount of nitrogen and phosphor than do plants.

Most of these non-photosynthesizing prokaryotes are specialized to consume organic matter, which is important in ecosystems like the Baltic Sea. Organic matter is released upon cell death of for example phytoplankton, but can also be flushed into the sea from land. These heterotrophic prokaryotes repackages the organic matter so that it can be propagated to higher trophic levels through grazing by larger plankton, like protozoa (Fenchel 2008). Furthermore, a final argument to convince someone about the importance of microbes, if one is ever needed, is that they produced oxygen for almost 2 billion years before land plants even came to exist (Falkowski 2012).

The Baltic Sea is scientifically interesting for several reasons. First of all, with gradients of salinity, oxygen, nutrients and temperature, it contains several vastly different but yet connected local environments to study. Furthermore, it is also the world's second largest basin of brackish water and thus also host for a brackish microbiome which is explored in paper III. Finally, it is subjected to a large deposit of nutrients from its surrounding land areas, causing eutrophication. The effects of the nutrient load are also worsened by the long retention time of the Baltic Sea water.

The following sections will introduce methods used to study environmental microbes but saving computational details for the next chapter.

Culturing

The gold standard for microbiological studies are based on isolation and culturing of cells, producing a clonal population. This enables experiments to be performed where the functional capabilities of the cells, as well as individual gene functions, can be investigated. Furthermore, the clonal population is also ideal for sequencing experiments. However, culturing is complicated for most microbial organisms due to differences in optimal growing conditions. Some organisms are also dependent on other members within their normal community in a complex pattern, further complicating a culturing approach.

Culturing of microbes is therefore often a very time consuming task, and for most environments, at least tens to hundreds of different species would have to be cultured in order to reach a reasonable coverage of the community in question. Even with a sufficient proportion of the present community available in culture, the proportion of different organisms present in different samples would still not be directly available, which leads us to the subject of metabarcoding.

Metabarcoding

A common method to identify and quantify organisms within a microbial community involves the study of the gene coding for the small subunit ribosomal RNA (rRNA). For bacteria and archaea, the gene used is the 16S rRNA gene while for eukaryotes it is the 18S rRNA gene. This gene is present in most cellular organisms and the structure of its sequence is particularly useful for this task. The method of acquiring the DNA sequence of this gene, or simply *sequencing* this gene, for members of the community, will here go under the name *metabarcoding*.

The first characterizations of the 16S and 18S genes actually used the resulting RNA product which is abundant in the cell. This kind of sequencing was used as early as 1977 to establish archaea as a group of organisms on the same level of independence as bacteria and eukaryotes (Woese and Fox 1977). The first characterization by sequencing of a microbial community focused on a section of the large subunit of ribosomal RNA (Stahl et al. 1985) but researchers shortly turned to the small subunit for the higher resolution it offered. Since then, the methods of metabarcoding have evolved and grown immensely popular. Most environment types have now at least partially been studied, including the Baltic Sea (Herlemann et al. 2011; Hu et al. 2016).

Metabarcoding can be said to answer the question ‘*who’s there?*’ and to give a good estimate of the relative abundance of the members in the community. However, from only metabarcoding studies, many of the organisms studied are not known to much further extent than by their 16S or 18S sequence. While the ecological role of a species might be hypothesized from the specifics of the samples where it was quantified, it cannot be verified or fully understood without further investigations. One of the absolutely best sources of information for the functional potential of a species is its genome, which contain many more genes than the single one studied by metabarcoding.

To acquire the DNA sequence for a single microbial species' genome, often called to *sequence* the genome, usually requires the isolation and cultivation of that species. However, due to the previously mentioned difficulties with culturing of most microbes, a regular genomics approach is not a feasible way to study a community of microbes. Furthermore, to sequence individual cells without culturing, so called *single cell sequencing*, is complicated and was not technically feasible until relatively recently (Zhang et al. 2006).

Metagenomics

Instead, to extend on the information available from metabarcoding: the answer to the question '*who's there?*', without needing to culture the organisms studied, researchers attempted to sequence any DNA fragment available in an environmental sample. This approach, which is called *metagenomics*, attempts to answer the question, '*what can they do?*', i.e. to determine the function of the community. To determine the function of a sequence is to functionally *annotate* the retrieved sequence. This is done by comparison to sequences available from cultured species and sequences which are sufficiently similar are assumed to have a similar function.

While metagenomics was possible using traditional low-throughput sequencing techniques, it blossomed with the advent of massive parallel sequencing. The increased throughput from the new machines and a decrease in cost of sequencing enabled a great number of large-scale metagenomics sequencing projects. Among the massive parallel sequencing technologies, the Illumina HiSeq machine deserves a special mention. It has, in different versions, been used for several large scale metagenomic projects and also for all papers included in this thesis.

The term 'metagenome' for the collective genome of a microflora was presented already in 1998 (Handelsman et al. 1998). However, using the current meaning of the word, the first metagenomic study of prokaryotes was published some years later (Tyson et al. 2004). This was coincidentally also the first successful application of metagenomic binning, which will be covered in the next section. This study sequenced a relatively simple community inhabiting a biofilm in acid mine drainage water. Despite the low-throughput sequencing technique used, this study not only managed to recover the genomes of the dominant species of the community, but also presented evidence for extensive homologous recombination between strains for one of these species. This evolutionary process was previously thought to be rare among prokaryotes. Another very early metagenomic study, which turned out to be ground-breaking for marine metagenomics, studied the Sargasso Sea (Venter et al. 2004). This study only used a low-throughput sequencing technique (Sanger sequencing). However, machines that were out of job after the human genome had been finished allowed for a massive scale, generating close to 2 million sequence reads. The vast diversity that this study displayed inspired several initiatives with cruises of the global oceans, collecting water samples for sequencing. All papers included in this thesis, except Paper II, can be said to be part of the field of marine metagenomics.

However, metagenomics applied to environmental samples have not attracted as much attention as studies of human associated microbiomes. At least two ambitious projects have tried to map the human microbiome in detail. The mainly European initiative MetaHit focused on the gut microbiome only (Qin et al. 2010), while the mainly North-American Human Microbiome Project studied a wide range of body sites (Human Microbiome Project Consortium 2012). Both of these projects aimed to build reference catalogues of sequences found within respective microbial community and were successful in doing so. A similar approach to construct a somewhat complete gene catalogue was applied in Paper I to create a reference assembly of Baltic Sea microbial communities.

Metagenomic binning

While metagenomics might be able to estimate the function, or at least the functional potential of the entire community, it is not clear what function is linked to which species. This leads us to the main focus of this thesis, metagenomic *binning*. Through metagenomic binning, metagenomic sequences which are believed to originate from the same species are placed together in a *bin*, without necessarily having any prior knowledge of the species. This enables the functional annotations of the sequences to be connected in a meaningful way. For example complete metabolic pathways can be reconstructed based on coexisting genes. If any of the sequences within a bin carry taxonomic information, taxonomic information can be connected to the functions.

The ability to extract genomes from environmental samples have greatly expanded our knowledge about the tree of life and led to important scientific discoveries. One of the most important of these is the discovery of a novel archaeal phylum with clear similarities to the eukaryotic domain, hypothesized to contain the ancestor to all eukaryotic organisms (Spang et al. 2015). As was previously described, the first metagenomic study also discovered the first clear evidence for homologous recombination within prokaryotes. Other large-scale studies have expanded the tree of life with hundreds (Brown et al. 2015) or thousands (Donovan H. Parks et al. 2017) of new species, respectively.

Furthermore, studies focusing on specific environments have recovered a substantial proportion of those environments' microbial communities. For example, several hundred genomes were recovered from a single large-scale study of human gut samples (Nielsen et al. 2014). Metagenomic binning can also be used to investigate specific biotechnological applications. A study of the somewhat exotic moose gut can serve as an example of this (Svartström et al. 2017). In this study, 99 genomes were reconstructed and a large proportion of these are believed to play a crucial role in the degradation of cellulose, an important biochemical process for a potential biofuel production.

Genomes have also been reconstructed for the ocean microbiomes of the world. From the global sailing cruise Tara Oceans, 92 metagenomic samples was processed, achieving 957 non-redundant genomes (Delmont et al. 2017). A more local study, focusing on the Baltic Sea, is presented in paper III, where 30 non-redundant genomes were recovered, followed up a tenfold expansion in paper IV Methods for binning metagenomic sequences will be presented with technological details later.

While the reconstruction of prokaryotic genomes is the main objective of metagenomic binning, some recent studies have also successfully reconstructed eukaryotic genomes (Delmont et al. 2017; West et al. 2018). Eukaryotic organisms are generally more complex than prokaryotic ones, and so are eukaryotic genomes. For at least three reasons, metagenomic binning of eukaryotic cells is more complicated. First of all, eukaryotic genomes are normally larger than prokaryotic ones. Secondly, many eukaryotic organisms have two copies (diploid) or more (polyploid) of every DNA-molecule with some variation between them. Lastly, and perhaps contradictory at first glance, is that eukaryotic genomes also contain regions of low complexity. These regions can contain short repetitive sequences which are difficult to distinguish from each other. All together, the reconstruction of eukaryotic genomes are not straight-forward even for data from a cultured species, let alone so from metagenomic data.

While metagenomic binning can place sequences from the same species together in a bin, the genomes of cells within the same species might be different. These cells are said to represent different strains. The presence of multiple strains from the same species pose a problem to not only binning but to all metagenomic analysis. The problem, and the solution, is furthermore dual, where some methods focus on identifying gene sets corresponding to each strain (Scholz et al. 2016) while others aim to identify the exact sequence for a strain (Luo et al. 2015; Truong et al. 2017; Nicholls et al. 2018). Another approach is called *strain resolved binning* which strives to refine binning results in order to find both the gene set and the exact sequence of the strain at once (Quince et al. 2017).

From medicine, it is known that one strain might be pathogenic even though other strains from the same species are not (Segata 2018). It is therefore reasonable to assume that environmental strains also differ widely in ecological function. Environmental studies using strain resolved binning are so far very sparse. It has, however, been applied to ocean water samples (Quince et al. 2017) where a connection between genome sizes and strain divergence was shown. It is my belief that methods with resolution down to the strain level will gain popularity in a close future.

Single cell sequencing

Alongside the development of metagenomic techniques, methods for directly sequencing individual cells, so called *single cell sequencing*, have evolved and matured. The main issue with single cell sequencing lies within the field of biochemistry. In order to conduct genome sequencing, a sufficient input quantity of DNA is required. Within a single cell, there is not enough DNA to fulfill this requirement which necessitates DNA amplification prior to sequencing. However, amplification is complicated when starting with only a single copy of each DNA molecule, as opposed to in metagenomics where multiple cells with close to identical molecules are assumed to be present. This often leads to uneven amplification where some regions are underrepresented or missing in the resulting sequencing output.

Furthermore, many ecological hypotheses require data for abundances for organisms over multiple samples in order to be tested. To acquire this through only single cell sequencing would require sequencing a very large number of cells. On the other hand, if a single cell sequencing approach would be combined with metagenomics, this number could probably be reduced drastically.

Without any pre-selection of cells, the vast majority of cells sequenced would originate from the most abundant species. This is the case also for metagenomics, but the marginal cost for adding individual cells is higher for single cell sequencing. Therefore, in order to sequence sufficient amounts of any less abundant species, pre-selection of which cells to sequence is often necessary. This includes screening of cell types using rRNA amplification and sequencing. All together, this makes single cell sequencing rather elaborate.

Phasing and long read sequencing

Related to strain resolution and single cell sequencing is the question of *phasing*. Phasing is a process of connecting sequence variants that are further apart than the sequencing machine normally can cover. This distance is dependent of the machine’s read length. With phasing, the sequence which correspond to a specific strain can be obtained. A distinction can be made between phasing based on bioinformatic methods, such as strain resolved binning previously discussed, and phasing performed by molecular methods. This section will focus on such molecular methods. Successful phasing should also simplify or perhaps even eliminate the need for metagenomic binning since much longer sequences can be constructed directly from the metagenomic sequencing data.

The most promising phasing methods need specific laboratory preparation of the DNA molecules which have not been performed for any of the samples included in the papers included here. The application of molecular phasing methods to metagenomics have shown some promising results (Bishara et al. 2018) and if combined with recent developments to decrease the cost (Redin et al. 2017) of phasing, the future looks bright for these methods.

The development of sequencing machines for so called *long-read sequencing* is continuously on going. These machines can produce several order of magnitudes longer reads than the commonly used Illumina HiSeq. While longer reads would be beneficial for most applications of metagenomics, these methods can currently not match either the accuracy or the price per base offered by Illumina sequencing machines.

Bioinformatic methods

There is no clearly established definition on what it means to be a *bioinformatician*. What someone intends with the word ranges from: a computer scientist who mainly develops new algorithms; a data scientist, statistician, or biologist that uses scripting for analysing data and evaluating hypotheses; or a system administrator who maintains computer software and sometimes hardware. To me, a bioinformatician is someone who does a little bit of all of these things, and irrespective of job title, intends to apply it within biology. In this chapter I will present bioinformatic methods, by which I mean software developed to solve a specific task within biology; in this case to be used within metagenomics.

Instead of trying to cover all aspects of metagenomic bioinformatics, a special focus will be on methods suitable for environments which are not well represented by reference databases. Therefore, methods heavily relying on these databases will not be covered. Furthermore, to allow a more in-depth coverage of metagenomic binning methods, several important methods are out of the scope for this thesis and will not be covered at all. These include methods for building phylogenetic trees, performing taxonomic assignment, constructing global alignments, and methods specifically designed for read-based metagenomic analysis. Methods that will be mentioned but not described in detail are those related to gene prediction and functional annotation. On the other hand, outside of the above mentioned definition of bioinformatic software lies several utilities which have been very important to me during my time as a phd student and which therefore yet deserves to be mentioned.

To some biologists the use of the bash command line is synonymous with bioinformatics. While intriguing at first, it offers an efficient and unifying interface to any unix computer. Especially when working with remote servers where graphical user interfaces are often not present, knowledge of the command line is key. When using remote servers, a terminal demultiplexer such as tmux will increase productivity. It will allow multiple command line sessions to remain active, even if for example your internet disconnects or when restarting your laptop.

Since bioinformaticians often work with plain text files, general tools developed for a broad use are ideal. Tools like *paste*, *cut* and *grep* follow one of the principles of unix systems: they can perform a single task, but really well. Another unix tool, *sed* - stream editor, does exactly what its name says: it modifies streams of characters. While these edits are extremely versatile, sed is especially used for search and replace actions on text files. Finally, other tools help out with installing software, e.g. *conda*, or manages custom computational workflows, e.g. *snakemake*.

Two methods which are not specific to bioinformatics but that are commonly used within the field are Principal Components Analysis (PCA) and Expectation Maximization (the EM-algorithm). These methods are both used by metagenomic binning methods and are therefore briefly described here. PCA is mathematically a linear transformation, commonly used to visualize a high-dimensional dataset. The transformation is constructed to map the highest variation in the original data along the first axis, the second most along the second and so on. Visualizations in two dimensions simply use the first two of these axes (or components) as x and y. However, it is also possible to decide on a given fraction of variance that is to be kept, and keep just enough of the first components to do so. This approach is common for the methods that I will describe later. In practice, this reduces the input data, which speeds up computations, without losing too much information.

A common method for statistical inference, when an exact solution is not easily obtainable, is the EM-algorithm. Despite its name, this is rather a collection of algorithms which is most often applied to clustering. More specifically, it is used for clustering where an explicit statistical distribution can be assumed for each cluster. The algorithm uses an iterative approach which is guaranteed to find at least a local maximum of the global likelihood of the model. It is applied by first reformulating the model so that cluster memberships are explicit numerical variables. Out of two repeated steps, the first is called the *expectation* step. In this step, the expected values of cluster memberships are calculated. These expectation values can be seen as fuzzy cluster memberships which are used in the second step, the *maximization* step. In this step, all other model parameters, such as the cluster characteristics are estimated using maximum likelihood, keeping the cluster memberships fixed. These two steps are repeated until convergence is achieved (Hastie, Tibshirani, and Friedman 2001).

Classic bioinformatic tools

While many of the methods that will be presented here have been developed fairly recently, some have been proven by time and form a foundation for much bioinformatic research. One of these algorithms is the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) which is used to find similar sequences within a large database of sequences for some given query sequence(s). The objective of BLAST is to find the subsequences with the best match to each given query sequence. Since the matching regions can be a small fraction of the query sequence (and the subject sequence), this is termed *local alignment*. To determine the best match a scoring scheme is used where, for DNA sequences, there are only the case of match and mismatch. For proteins, a more biologically informed scoring scheme is used. Since it was developed the amount of sequences available in databases have grown immensely and several more efficient solutions have been proposed over the years, yet BLAST remains the de facto standard for sequence database queries.

The basic BLAST algorithm does not take into account that some parts of a sequence are highly conserved while others are less so. However, it makes sense that mismatches within conserved regions are much less probable and should affect the scoring of the alignment more than mismatches in other regions. This fact is utilised in Hidden Markov Model (HMM) profiles, which are tightly connected with the software HMMER, perhaps the most used implementation of HMM profiles (Eddy 1998). An HMM profile is created from a multiple sequence alignment of for example, members of a protein family. From this alignment, a probabilistic model is created governing how a protein sequence could be generated from that profile. This generation is only conceptual. The model is used for scoring purposes: the probability that a given sequence would be generated from the model is interpreted as a score for how likely the sequence is to belong to that, for example, protein family.

Profile searches are especially useful to achieve high sensitivity when searching against databases of groups of related sequences, so called orthologous groups. One such database, launched already in 1997, is the database for Clusters of Orthologous Groups (COGs). The COG database was constructed using pairwise alignments and reciprocal best hits of all genes available from five distant lineages represented by seven species. Furthermore, the COG database offers manually curated names and information for individual COGs as far as possible. To match any given gene against the COG database, a commonly used software is RPSBLAST, which is included in the BLAST suite of softwares. RPSBLAST is constructed to use protein profiles, however, these profiles are different to HMM profiles as they are not based on hidden Markov models. RPSBLAST was used in paper II and paper III to match genes against the COG database.

While the manual curation of COG annotations has clear advantages, it requires a major effort from the scientists maintaining the database. This could explain why no major update of the COG database has been released since 2003. Another database initiative using automated creation of orthologous groups is the eggNOG database (evolutionary genealogy of genes: Non-supervised Orthologous Groups). It consists of 1.9 million HMM profiles hierarchically structured according to taxonomy (Huerta-Cepas et al. 2016). The eggNOG database includes all COGs as a subset, placed on the top taxonomic level together with just under 200,000 other groups. Furthermore, all groups within eggNOG are automatically annotated based on all annotation information available for the underlying individual protein sequences. The eggNOG database was used in paper I and paper V to find gene homologs for the obtained sequences.

Here, I have described two methods to perform local alignment or profile searches with high sensitivity. Another subset of tools for alignment are those which are optimized for precision rather than sensitivity. An application where this is a good choice is when short sequencing reads are to be aligned against a closely related reference genome, i.e. from the same species. This process is often called to *map* the reads, and hence the final piece of classic bioinformatic softwares are *read mappers*. I will not point out a single best tool for mapping reads since there are several comparable implementations. However, with modern sequencing technologies that produces a large number of overlapping reads, the computational efficiency of a read mapper is key. All of the most commonly used implementations are based on the clever Burrows-Wheeler transform (BWT) which enables fast lookup of exact matches with a small memory footprint. Due to sequencing errors and/or biological variants, finding only exact matches is typically not sufficient. One approach to find inexact matches, implemented in the Bowtie2 software, uses short substrings of the reads (Langmead and Salzberg 2012). Exact matches for these substrings are identified using the BWT and are extended to find a good match for the whole read. Bowtie2 is used for read mapping in Paper I-IV.

K-mer based methods

In this section additional bioinformatic tools which are useful for metagenomic binning will be presented. A unifying feature of these methods is that they are all based on *k*-mers. BLAST and BWT algorithms (in some implementations) also uses *k*-mers, although inexplicitly. A *k*-mer is a substring of length *k* from a given sequence, generated as shown in Figure 1. The length *k* varies from application to application.

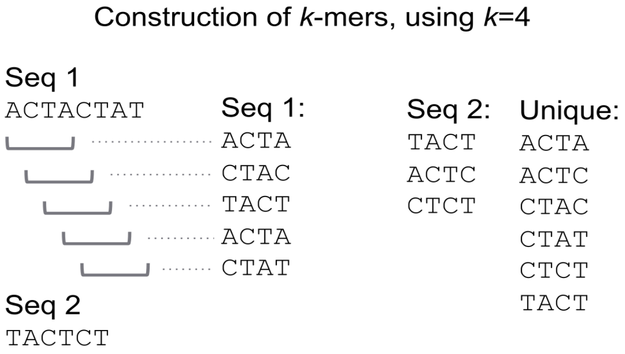

Figure 1: Construction of *k*-mers from two short DNA sequences. *K*-mers are constructed using a “sliding window” of size *k*.

Perhaps the most complex application of *k*-mers that will be covered here is that within *assembly*. Assembly is the process of constructing longer sequences using the reads from a sequencing machine. These longer sequences are called contigs, after the word contiguous. This thesis is dedicated to metagenomic analysis based on contigs, as opposed to read-based analysis. The latter tends to be more dependent on comparisons against databases of reference sequences and therefore less suitable for studies of environments not yet well covered by these databases. However, these databases can be extended using contigs constructed from metagenomic sequencing and hence enable future read-based analysis. Assembly is a fundamental step for the reconstruction of genomes from metagenomes, the main focus of this thesis.

Successful assembly depends on properly prepared DNA molecules and sufficient sequencing depth. This ensures that the sequencing reads overlap in such a way that it is possible to form longer consensus sequences. Simply comparing all reads against each other to find overlaps between them quickly becomes too computationally heavy. This is where *k*-mers come into play through the creation of a *de Bruijn* graph.

A de Bruijn graph is a data structure built by connected *k*-mers. Two *k*-mers are connected in the graph if they appear consecutively in any read. This data structure only store the reads represented as *k*-mers and does not store information on from which read each *k*-mer originated. Representing the reads in this way can save memory since the same *k*-mer is often found in many reads. However, the main advantage is that constructing consensus sequences from reads translates to finding paths within the de Bruijn graph. The assembly program used for paper II and paper III is called Ray and is used due to its highly parallel implementation, enabling the use of multiple server computers simultaneously. For papers I and IV, an assembly program called Megahit was instead used (Li et al. 2015), taking advantage of a different strategy, presented below.

Most assembly programs allows the user to choose the value of *k* to use. Short *k*-mers allow for smaller overlaps between reads to result in contigs. On the other hand, short *k*-mers are more likely to exist in multiple locations on a single genome, or even on multiple genomes, and therefore be present in different reads which are not supposed to be assembled together. In general, the higher sequencing depth and longer reads obtained, the longer *k*-mer is possible to use. However, for most metagenomic samples, some species will be present in low abundance and hence obtain a lower sequencing depth. These species will therefore not be assembled well when using a larger value for *k*. In paper III this problem was solved by running the Ray assembler several times with different values for *k*. The resulting contigs were then merged by explicitly searching for overlaps between them. In this way the benefits of both short and long *k*-mers were obtained.

The assembly program used for paper I and paper V, Megahit, have instead directly implemented a multiple *k*-mer approach (Li et al. 2015). In its implementation, Megahit builds the new de Bruijn graph by *k*-mers from the reads *and* from the contigs created from the previous step, if any. Megahit uses iteratively larger values for *k* for each step. The default range starts with *k*=21 and eight steps up to the maximum *k*=141. Of course *k*=141 does not generate any graph if built only on reads that are shorter than 141, for example when using 125 bp reads. However, since Megahit also includes the contigs from the previous step when building the graph, this will work. It should be noted though that running Megahit with a maximum *k* larger than the maximum read length does not improve the assembly. Furthermore, Megahit uses a very memory efficient implementation of the de Bruijn graph. This makes it possible to assemble most metagenomic samples on a single, fairly standard, server.

In assembly based studies, the choice is often between assembling all available samples together or creating individual assemblies per sample. In Paper III and V individual-sample assemblies were used to avoid mixing sequences between related strains more than necessary. With this strategy dominant strains are expected to assemble consistently since the complexity of the individual sample is lower than when all samples are combined. Less abundant strains in one sample might be more highly abundant in a different sample and thus, focusing on only dominant strains for each sample might not be as wasteful of data as one might think. However, some strains can be low in abundance in all of the available samples. If these are to be assembled properly a co-assembly approach might be more appropriate, which combines all or at least several samples prior to assembly. A co-assembly approach was used in Papers I-II. The specific implications to metagenomic binning of these approaches will be discussed further in the next chapter.

The next *k*-mer based tool to be covered is related to the Burrows-Wheeler Transform short read aligners. One of the most common use-case for short read aligners within metagenomics is to quantify the abundance of the contigs constructed by the assembler. However, it is not strictly necessary to align all reads against the contigs in order to quantify them. This is the idea behind the tool Kallisto which thereby is able to achieve faster execution time compared to regular short read aligners (Bray et al. 2016).

Much like the assemblers previously covered, Kallisto is also based on a de Bruijn graph, but built using the reference sequences directly and not from the reads. Furthermore, instead of a regular de Bruijn graph the data structure used by Kallisto also keeps track of which reference sequence each *k*-mer originated from. Kallisto then introduces the concept of pseudoalignment where a read is only identified with a reference sequence without identifying the exact position within that sequence. Kallisto identifies reference sequences that matches all *k*-mers within the read. The quantification of the reference sequence is however determined by the EM-algorithm, iterating over the assignments and adjusting counts per reference sequence to find an estimate of the most likely counts. For example, if the *k*-mers of a read matches several different reference sequences the likelihood is maximized if the read is placed on the reference sequence that already have the most reads assigned to it. It should be noted that Kallisto was not developed for metagenomics but for quantification of transcripts for RNA-sequencing. Therefore, the validity of using Kallisto within metagenomics was evaluated and the result is shown in Figure 3. As can be seen, the coverage values are highly correlated between the two methods: Kallisto and the traditionally used Bowtie2. Out of these two the benefit of Kallisto is its very quick run time. However, Bowtie2 offers additional information since the exact placing of the reads allows inspection for e.g. patterns of mismatched bases reflecting genetic variation. Bowtie2 was used in Papers I-III and Kallisto was used for quantification of metagenomic contigs in paper V.

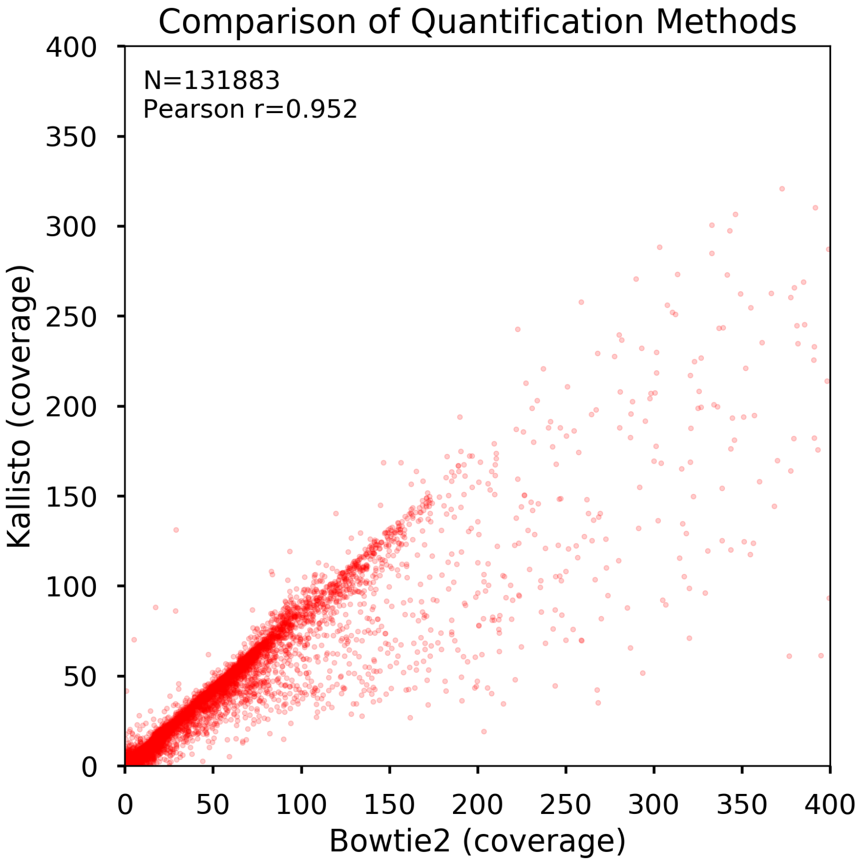

Figure 2: Quantification comparison between Kallisto and Bowtie2. The sample containing the reads was not the same as the sample that generated the contigs. It should be noted that the vast majority of dots (N=101122) are located within the square where both values are smaller than 0.5. The largest values (N=457) are not shown.

The next application of *k*-mers within bioinformatics to be discussed is that within MinHash-based algorithms. These bioinformatic algorithms are all fairly recent. They are used to give extremely fast but approximate average nucleotide identity values between two sequences. To achieve this approximate value for the nucleotide identity, two sequences are compared by only comparing a subset of their *k*-mers. However, if the selection of the subsets would be random the variance of the estimate would be very high for small subsets. MinHash is a clever way of producing these subsets so that the variance of the estimate remains low.

The foundation of the MinHash algorithm is hash functions which in turn are fundamental computational methods to map arbitrary 'objects' to integers. In our case these objects will be *k*-mers. Correctly constructed the integers produced by the hash functions enforces an arbitrary but reproducible ordering of *k*-mers. Then based on this ordering the *l* first of these are selected for each sequence, where *l* is the chosen size of the subsets. This allows very small subsets to be used with an acceptable precision of the estimate. One implementation of MinHash for bioinformatics is Mash (Ondov et al. 2016), which was used in paper IV to compare genomes obtained with two different methods. Mash was also used through a wrapper called FastANI in paper V to cluster a large set of genomes into groups corresponding to the species level.

The rest of this *k*-mer focused section will connect to the following chapter where we will discuss metagenomic binning. In this application a different aspect of *k*-mers is used. Namely that small values for *k* can give *k*-mers unspecific enough to match several positions on a sequence. This was previously discussed as a negative thing, e.g. in the context of assembly. Here it will instead be a positive thing where individual *k*-mers are assumed to be found in many positions within the same genome.

The first example of this usage of *k*-mers that will be presented is found within the gene prediction software Prodigal (Hyatt et al. 2010). Prodigal is an extremely fast gene predictor capable of finding genes on both genomes and metagenomic sequences without any additional information. It uses a mixture of knowledge acquired from manual curation of genomes together with a large set of parameters which are trained on each genome or individual sequence. Among several other metrics it uses *k*-mers with *k*=6 to score different gene models. For example, if two suggested sets of genes are weighted against each other, Prodigal would (among other things) compare the 6-mer usage within suggested genes compared to the entire sequence. The gene set with a more specific 6-mer profile is then believed to be the most likely out of the two. In this sense the 6-mers are only supposed to be specific enough to, on average, distinguish gene content from intergenic regions.

Sufficiently short *k*-mers have similarly been shown to carry a phylogenetic signal. This signal is furthermore somewhat consistent over different regions of the genome (Dick et al. 2009). For example, a method called EukRep was recently shown to be able to distinguish between eukaryotic sequences and sequences of prokaryotic origin within a metagenome, only based on patterns of *k*-mers (using *k*=5) (West et al. 2018). This method was applied in paper V.

For *k*=4, which is the most established choice within metagenomic binning, only 256 possible *k*-mers exists. This is often reduced further by considering two *k*-mers identical if they are the reverse complement of each other. This allows 4-mers to be general enough that each individual *k*-mer is to be found within most sequences. A *k*-mer profile of a sequence, commonly called the *nucleotide composition* of the sequence, is constructed by counting all *k*-mers present in the sequence. These counts are then normalized by the total number of *k*-mers in the sequence. Two different sequences can then easily be compared by the similarity of their nucleotide composition. The idea behind this is slightly counterintuitive: Why would two sequences have similar nucleotide composition just because they originate from the same genome, even if they originate from different parts of that genome? No single explanation for this has been widely accepted. It could be due to mutational bias, allowing different species to differentiate in a somewhat regular manner. However, to some extent this has been observed to be true. Therefore, the use of *k*=4 have a long tradition within metagenomic binning, the subject for the next chapter.

Metagenomic binning

As was mentioned previously metagenomic binning is the main focus of this thesis and it will be the only focus of this chapter. The chapter will start with some background before continuing with a detailed description of published binning methods and a small performance comparison of these methods. Finally, useful tools surrounding the actual task of metagenomic binning will be presented. In paper II a method for metagenomic binning named CONCOCT is presented. For completeness CONCOCT will also be be briefly presented in this chapter.

In order to recover genomes from metagenomes, binning of contigs is necessary. This is because the length of the contigs which are output from the assembly process are typically short. Contigs are very rarely longer than 100 kilobases, and often much shorter, while the genomes of most free-living organisms are at least one order of magnitude larger. While methods exist to improve the assembly further, the most effective ones require specific laboratory treatment prior to sequencing, not commonly applied. When describing available methods to perform binning I will only focus on automatic methods since manually curated approaches to a large extent depend on the user. However, to give a historical background, genomes manually reconstructed from metagenomes will also be considered. Furthermore, there is a distinction between supervised and unsupervised methods. Supervised binning methods, to some extent, use available data from public databases. Some methods can be said to be semi-supervised, meaning it only partially depends on reference data.

The first genomes to be reconstructed from metagenomic data originated from samples from acid mine drainage water. The low microbial diversity of this hostile environment allowed genomes to be manually recovered using a combination of G+C content and sequencing depth, *coverage* (Tyson et al. 2004). Shortly after this, a method based on so called Self-Organizing Maps (SOM) was published. This method transforms tetranucleotide frequencies into a two dimensional space (the map) where clusters could be identified (Abe et al. 2005). This method is not automatic since the clusters are located manually on the map by the user. The program CompostBin introduced a semi-supervised algorithm. It does not need training based on reference genomes, but uses phylogenetic marker genes found on input sequences (Chatterji, Yamazaki, and Bai 2008). The first automatic and completely unsupervised method was LikelyBin (Kislyuk et al. 2009), that clusters contigs by nucleotide composition using a probabilistic model.

Further developments to metagenomic binning methods were however necessary since nucleotide composition has a limited resolution. The next major step in the development of these methods was to reintroduce sequencing coverage as a source of information. The argument for using sequencing coverage is as follows: fragments that originate from the same genome should be present in equal amounts in the sample and sequencing coverage is an approximate measurement of fragment abundance. Hence, sequences originating from the same genome should have similar sequencing coverage values. However, by chance, two different genomes can have equal abundances in a sample and therefore be impossible to separate using only coverage for this sample. If several samples is used the chance of identical abundance in all of the samples is however very small.

The effectiveness of binning using multiple samples was first shown by simply plotting the coverage values for the two samples in a scatter plot and colour the dots according to G+C content (Albertsen et al. 2013). The clusters were further refined using PCA built on tetranucleotide frequencies. This manual approach was shown to improve the results achieved by previous methods. However, manual methods rely heavily on a skilled user, not always available, which is why automatic methods are often preferable.

A large number of methods for automatic binning have since been published. Some of these will be described in the following parts of this chapter. The differences between them can often be technical and non-trivial. Therefore, following this description, a simple performance comparison between the described tools will be presented. However, before continuing this chapter with descriptions of individual tools, the question whether to use individual-sample assemblies or a co-assembly will be addressed.

As was described in the previous chapter a co-assembly is beneficial for species which would otherwise not reach a sufficient sequencing depth. Furthermore, co-assembly is perhaps also more theoretically pleasant for the type of read alignment performed in modern metagenomic binning. The reason for this is most easily explained by looking at the opposite alternative. When binning is performed on individual-sample assemblies, all read files are aligned against each assembly. If the species from which a read truly originates from is not present within that specific assembly, the read might be aligned against a contig belonging to a different species. This should affect the binning results negatively. On the other hand, if a co-assembly strategy is used, all reads have been used to construct the assembly and this should be less of a problem. In practice, however, binning results from individual assemblies have been shown more successful than the corresponding results from a co-assembly. This was found in a comparison conducted by us leading up to Paper III, and has also been studied in detail later (Olm et al. 2017). In this detailed study, individual-sample assembly approaches not only produced more high-quality genomes but these had also longer contigs and were estimated to be more complete than those produced form a co-assembly. In Paper II, a co-assembly based strategy was used to perform metagenomic binning. In Paper III and V, this approach was modified to perform binning on individual assemblies from the Baltic Sea.

Canopy

One of the first and arguably simplest methods that use coverage over multiple samples is Canopy (Nielsen et al. 2014). This method actually clusters genes extracted from contigs and not the actual contigs. Genes are clustered using Pearson correlation for the coverage patterns and only includes tetranucleotide frequencies as an optional quality screening step. The genes found correlating form putative clusters which are then filtered in two consecutive steps. A cluster resulting from each of these steps are respectively named a canopy, a co-abundance gene group (CAG) and a metagenomic species (MGS).

In detail, Canopy clustering starts by choosing a seed gene randomly and then screen all other genes, recruiting all those with a correlation coefficient of at least 0.9 to form a canopy. This search is repeated iteratively using the median coverage pattern to compare against all other genes. The iteration continues for this single canopy until the median stabilizes. New canopies are formed in the same way until all genes have been assigned. These clusters are then filtered so that the approved ones, CAGs, contain at least three genes and have a non-zero coverage in at least four samples. These rejection criterias are, however, all possible to adjust. To approve a CAG as an MGS and thereby assigning it as a putative genome, the CAG is required to contain at least 700 genes.

The fact that Canopy uses genes instead of contigs could be seen as both a strength and a weakness. It allows the detection of strain specific gene sets since genes are seen as individual entities. This often leads to non-core gene sets to be placed in separate clusters. On the other hand, connecting those clusters with the MGS corresponding to the core gene set usually has to be done in an ad-hoc fashion. As an example, in the original Canopy paper, most of the identified antibiotic resistance genes were not located within a CAG. It was argued that this is consistent with what is expected, since most such genes were known to "act alone". However, the fact that two genes are located on the same contig is a very strong indication that these two genes originate from the same genome. Simply ignoring this information should reduce the efficiency of metagenomic binning.

MetaBAT

A method for performing binning with claims of both speed and accuracy is MetaBAT (Kang et al. 2015), which uses both tetranucleotide information and coverage over multiple samples. When designing the program, a distance metric based on tetranucleotide information was derived from comparisons of a large amount of contig pairs of intra- or inter-species origins. In this empirical comparison the size of the contigs were also varied. It was found that distances between contigs shorter than 2000 bases are much noisier, why contigs shorter than this was not recommended to use for clustering. However, the minimum length of contigs that is possible to use for MetaBAT is 1500 bases. The distribution of coverage values for contigs known to originate from the same genome was also empirically investigated. This was done by downloading data from sequenced isolates and it was found that the distribution of coverage values could be best described using a normal distribution.

Using the distance metrics derived from tetranucleotide information and from coverage values, the algorithm constructs a matrix of pairwise distances between all contigs in the input data. For one contig at a time, starting with the contig with the highest coverage, the algorithm then assigns all other contigs within a fixed distance of the current contig to the same cluster. A *medoid* is defined as the contig within the cluster with the smallest average distance to the other contigs within the cluster. The algorithm then repeats the clustering steps, collecting all contigs within a fixed distance to the medoid and updating the medoid. If there are no updates to the medoid, the algorithm continues with a contig which have not been assigned to any cluster, again choosing the one with the highest coverage among the remaining contigs. By default, only sufficiently large clusters (>200kb) are reported, but as an optional step, unassigned contigs can be recruited to clusters based on the coverage information.

It is not entirely clear how storing the pairwise distance between all pairs of contigs can be so memory efficient. Despite this, MetaBAT is one of the most computationally efficient metagenomic binning algorithm available. This efficiency is perhaps a major reason why MetaBAT remains a popular choice when dealing with large datasets.

GroopM

One of the earliest algorithms to use coverage values over multiple samples was GroopM (Imelfort et al. 2014). This rather easy-to-use program has very complicated internals. The description of the algorithm presented here will therefore merely scratch the surface of the complete picture. Its first step is to load the coverage information from the read alignment files into a high dimensional space. It then continues by performing a carefully designed transformation of the coverage data to a three dimensional space. The information deduced from coverage is complemented with the tetranucleotide frequencies which are transformed using PCA, keeping at least 80% of the variance. The first clustering step uses a subset of contigs, forming so called preliminary bins. Starting with contigs located within the most contig-dense region in the transformed coverage space, clusters are formed according to similarity of both coverage, tetranucleotide patterns and contig lengths. This first step of binning is designed to be strict to avoid grouping of contigs from different genomes. Single genomes divided on multiple preliminary clusters is instead dealt in a subsequent step where sufficiently similar bins are merged. Bins are also checked for high within-cluster GC variation which is considered an indication of a chimeric bin. By default, GroopM only bins contigs longer than 1500 bp, but this is an adjustable parameter.

This was obviously a very brief description of GroopM. A somewhat complete description of the GroopM algorithm would fill several more pages of this thesis. It uses a wide range of machine learning techniques such as PCA, SOM, Gaussian Blur, K-nearest neighbour, and Hough transformation, all coupled together with heuristics and novel algorithms. One assumption used by GroopM which deserves a special mention is that it assumes contigs within a bin should have similar contig lengths. This assumption is not mentioned in the main text of the paper, but is clearly stated in the supplementary description of the algorithm. This assumption is not uncontroversial since the contig length depends on several factors. The average coverage of the contig is definitely one such factor which is already assumed to be similar within a bin. However, another factor which highly affects the length of contigs is the level of conservation for different regions of the genome. It is a very strong and most likely false statement to say that the level of conservation is more or less constant over different regions of a genome.

MaxBin

A relative straightforward probabilistic model is defined by the program MaxBin (Wu et al. 2014). Here, the tetranucleotide distances are assumed to originate from either of two normal distributions, inter- or intra-species. The parameters for these distributions were estimated empirically using a million pairs of contigs, simulated by extracting random subsequences from a large database of sequenced genomes. Even though the distributions did not appear normally distributed, this assumption was still made, motivated by the shape of the histograms and the fact that they were sufficiently separated. The coverage information for contigs is included by assuming a Poisson distribution. Version 2 of the program integrated coverage information over multiple samples into the algorithm which was previously intended to cluster a single metagenomic sample (Wu, Simmons, and Singer 2016).

The program starts by estimating the number of clusters by using a set of 107 so-called single-copy genes which are estimated to be present in each genome exactly once. With the estimated number of genomes, MaxBin applies a version of the EM-algorithm but with the added feature that after convergence the bins are checked again for single copy gene presence within the contigs. If a bin is found to have a median number of these genes above 2, it will be split by running it through the EM-algorithm again. MaxBin does not necessarily cluster all input contigs. If the probability, as defined by the model, that a contig belongs to the cluster it is assigned to is too low the contig will be discarded. Using prokaryotic single-copy genes to decide the number of bins is a clever trick to keep the clustering algorithm simple. However, this introduces a phylogenetic dependency making the algorithm less fit to cluster sequences originating from eukaryotes or viruses. Even prokaryotic plasmids which are likely to have a different coverage pattern than its hosts are unlikely to be clustered properly.

COCACOLA

The program COCACOLA (binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge) offers a mathematical non-probabilistic formulation of the clustering problem where an objective function is to be minimized (Lu et al. 2017). The authors conclude, however, that the exact formulation of the problem is NP-hard and make a reformulation corresponding to the soft clustering of the EM-algorithm, where contigs are not restricted to belong to exactly one cluster.

The model also contain a general way of including additional information to the model in the form of a network, where any other type of evidence that contigs belong to the same cluster can be entered. The suggested usage for this is to use paired-end read linkage and alignment against reference genomes. However, in their comparison the former only showed a marginally positive effect on the clustering performance and the latter introduces a dependency on what is present in reference databases.

The number of clusters is determined automatically but an initial guess is needed. The paper suggests a method which is to run k-means clustering until at least half of the clusters end up empty. The program also has a built in option to estimate the number of clusters from the presence of single copy genes in the input contigs. In their comparisons they show computational performance even better than MetaBAT.

ABAWACA

The software ABAWACA is not yet presented in a dedicated publication but is described in a paragraph in the first study where it was used (Brown et al. 2015). From this paragraph it is described to use a combination of nucleotide composition for three different values of k (k=1,k=2 and k=3) and coverage over multiple samples. Before clustering all contigs are split into 5 kb fragments. Whether or not two fragments originating from the same contig are clustered together or apart is used throughout the clustering as a quality estimate. The actual clustering starts with all contigs present in one single bin which is iteratively split into smaller parts in a hierarchical fashion. Each individual split is done based on a single dimension of the input data. This dimension is chosen as the one where the best split is obtained, as evaluated by the distribution of 5 kb fragments in relation to their original contig. However, when performing the actual split fragments from the same contig are kept within the same bin according to the majority vote. Both bins are required to have at least 50 fragments in order to be approved.

Using a strict cutoff for the minimum number of fragments to be contained within a bin is likely to be less successful for binning mobile elements or viruses. Furthermore, using a single dimension to separate contigs into two different bins might be computationally efficient but should also negatively affect precision in some cases.

MyCC

The program MyCC uses a combination of marker genes, nucleotide composition and coverage over multiple samples (Lin and Liao 2016). The used marker genes are 40 genes which are estimated to be universal within prokaryotes. Screening the input contigs for these genes can help assign phylogenetic information to the contigs. MyCC uses this information to refine the clustering results by either split or join preliminary clusters. MyCC uses a data transformation called t-SNE, commonly used within RNA-seq analysis.

Much like PCA, t-SNE can transform high-dimensional data to a space of, for example, two dimensions. While t-SNE give rise to visualizations in two dimensions which often resembles clusters, the validity of these can be questioned since the t-SNE transformation does not conserve distance between data points. Furthermore, t-SNE is somewhat parameter-dependent where in extreme cases, clusters can be observed in the transformed space where there are none in the original space (Wattenberg, Viégas, and Johnson 2016).

MyCC performs t-SNE transformation using all dimensions from nucleotide composition and coverage. However, in order to save memory and computational time, it only uses a subset of all contigs. The contigs not included in the first round of clustering is then assigned to pre-defined clusters from the first round. The clustering is performed using a method called affinity propagation which is finally corrected with the help of marker genes.

CONCOCT

The program presented in Paper III, which is also presented here for completeness, is called CONCOCT. It was early in its adoption of coverage over multiple samples to form an automatic clustering algorithm. Besides the coverage information, CONCOCT also uses tetranucleotide frequencies. Both of these types of information are normalized and merged into a shared matrix. The number of dimensions are reduced before applying the clustering algorithm using PCA, keeping at least 90% of the variance.

The clustering algorithm models each bin with a multivariate Gaussian distribution. These distributions are combined into a united model forming a so-called mixture model. A standard mixture model uses a fixed number of clusters. However, a method for clustering metagenomic contigs should ideally be flexible in this regard. CONCOCT uses a complex statistical method called a variational Bayesian approach to decrease the number of clusters from an initially large number. This initial number of clusters is recommended to be at least 2 to 3 times higher than number of clusters expected. It is a parameter that can be set by the user but the default value of 400 is in general a good choice. In an optional step clusters are evaluated on completeness and contamination using a custom script evaluating the presence of 36 single-copy COGs.

The performance of CONCOCT was evaluated on two simulated datasets and two real datasets. It was found to successfully separate clusters down to species level but to be less successful in separating strains from the same species. In the publication the running times of CONCOCT for the smaller simulated datasets were quick (around 4 minutes and 37 minutes respectively). In contrast, clustering of the largest of the real datasets took almost 36 hours to complete. This non-optimal scaling is still present in a recently available update of CONCOCT, but using a much higher degree of parallelization the absolute running times can often be reduced significantly. Furthermore, instead of running the actual clustering exactly 10 times and output the best of these the new version only run the clustering once, resulting in a 10 times speed-up, only marginally reducing the binning performance.

Evaluation of binning tools

A recent evaluation of common bioinformatic methods within metagenomics also included evaluation of metagenomic binning (Sczyrba et al. 2017). This evaluation could have pointed potential users in the right direction, but by design only included 1, 2 and 5 samples in the three different data sets respectively. The number of samples that are included within a single metagenomic binning study is of great importance. This is because, the more samples that are included, the higher the chance of including a sample where two genomes have clearly different abundance. Furthermore, real data sets, by design, often include many more samples than 5. Hence, the performance of methods in this comparison might not be representative to a real world data set.

This section will present a different comparison of the previously mentioned tools. The tool Canopy is not included since it is designed to bin genes and not contigs. Two different approaches are common when evaluating performance of binning tools. The first one uses a simulated dataset where the ground truth is know for all or at least most of the contigs. This enables computation of clustering metrics such as precision or recall which can be used to compare the tools' performances. However, constructing realistic simulated datasets is very hard. Real datasets often have high diversity and can also include fragments from genomes of low abundance. Furthermore, real datasets can include eukaryotic, prokaryotic and viral sequences, while simulated datasets rarely include more than prokaryotic sequences. For this reason, this comparison was performed on a real dataset obtained from Baltic Sea surface water.

Since the ground truth for the clustering problem is not known for a real dataset, the evaluation is instead based on the number of bins which are approved according to certain criteria. Two commonly used criteria are based on completeness and contamination. These parameters are efficiently estimated for each bin individually by CheckM, a tool which will be presented in more detail in the next section. In this comparison, the minimum required completeness was 70% and the maximum allowed contamination was 5%. The levels were chosen to match the controlled vocabulary of draft genome quality and correspond to substantially (70-90%) or nearly (>90%) complete with low (<5%) contamination. In this comparison, only one sample was binned and as with any comparison of this limited size, the results should not be interpreted as representative for any of these tools. It is however interesting to see how variable the performance is between these tools which in some sense can be described as relatively similar.

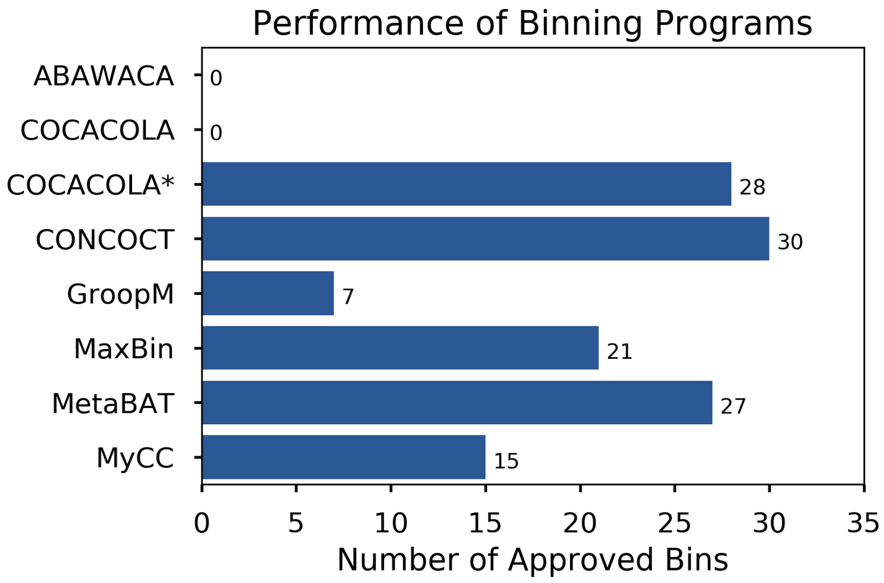

Figure 3: Evaluation of published programs for metagenomic binning. A single sample from the Baltic Sea served as input, but was quantified using in total 86 metagenomic samples from the Baltic Sea. Bins were approved using CheckM, requiring at least 70% completeness and maximum 5% estimated contamination. All programs were run using default parameters except COCACOLA* where the number of initial clusters were manually set to 200. All programs except COCACOLA* and MyCC was run by Sebastian Allard, Maja Andreasson, Saad Saeed and Cecilia Valdna Juhlin as part of their Bachelor thesis “The Bacterial Genome Puzzle”, KTH 2017.

The results of the comparison is shown in Figure 3. In summary, CONCOCT, MetaBAT and MaxBin performs well, as well as the second run of COCACOLA where the number of clusters were set manually. This was done since most clusters were highly contaminated in the first run of COCACOLA where the number of initial clusters were estimated by presence of single copy genes. To be fair, similar efforts would likely have improved the results of other tools which rely on single-copy genes to estimate number of clusters, in this case MyCC and MaxBin. Furthermore, the poor performance of some tools in this evaluation could be due to the use of pre-cut contigs. Contigs were cut into 10 Kb pieces as a part of the recommended workflow for CONCOCT, and since this is done prior to the time consuming quantification of contigs, it was kept as such for all tools. Especially the performance of ABAWACA likely suffers from this since it uses co-clustering of fragments from longer contigs as a fundamental metric. Shorter contigs reduces the number of fragments usable for such strategy. The program CONCOCT is described in detail in Paper II.

Other tools useful for binning

Several tasks are related to the actual task of clustering contigs into bins. This section will be dedicated to such tasks which include evaluating the completeness and contamination of bins and taxonomic classification of bins. These tasks were usually performed using custom scripts as they were in Paper II. However, there now exists several well maintained tools which greatly simplifies any metagenomic binning project.

The tool CheckM (Donovan H. Parks et al., n.d.) have rightfully been very successful. It was designed to evaluate bins in terms of completeness and contamination. As had been done before, this is achieved with CheckM by using single-copy genes. The principle is that the completeness is estimated by occurence of these genes and contamination is estimated by any observation of multiple copies of any such gene. Since different species can be evolutionary very distant, there only exists a few such genes valid to use for all prokaryotes. However, the precision in this type of estimates would benefit from having a larger number of such genes. This is achieved in CheckM by applying lineage-specific sets of single-copy genes. Therefore, a bin is evaluated by CheckM by first automatically determining its most likely lineage and then assessing its completeness and contamination using the set of single-copy genes associated with this lineage. The bins produced by a clustering method can be filtered using CheckM, only keeping sufficiently complete and uncontaminated ones. The results given by CheckM can in this way also be used to evaluate the performance of different binning methods, as was done in the previous section, without the use of reference genomes.

A similar method to CheckM which was originally developed for de novo sequencing of genomes is BUSCO. As opposed to CheckM, BUSCO is not explicitly designed for metagenomic binning and cannot automatically estimate the most suitable set of single-copy genes. However, BUSCO has one advantage to CheckM in that it can be used for eukaryotic genomes as well. CheckM, on the other hand, is limited to prokaryotes. Another tool which is useful for binning of eukaryotes is EukRep which was mentioned in the previous chapter. Both BUSCO and EukRep were used to obtain eukaryotic genomes in Paper V.

Although taxonomic annotation is outside of the scope for this thesis, I cannot resist mentioning a recently released tool which greatly aided the post-processing of produced bins in Paper V. This tool, which produces detailed taxonomic annotation of produced bins is the Genome Taxonomy DataBase ToolKit (gtdbtk). This tool is tightly connected to a recent preprint where a new taxonomy based on phylogenetic distance was suggested (D. H. Parks et al. 2018). The gtdbtk tool can place any genome or bin on this phylogenetic tree and assign a taxonomy based on its position. Other tools offer the same functionality. For example, Phylophlan was used in Paper III to assign taxonomy for individual bins (Segata et al. 2013). However, in order to include uncultured genomes, these had to be manually added to the database. With gtdbtk, the included database is supposed to be updated regularly and was built using all available genomes, even approved metagenomic bins. Furthermore, since the taxonomy is updated according to the phylogeny, even genomes within clades where no cultured genome exists can get detailed taxonomy. Taken together, gtdbtk is easy to use and enables detailed taxonomic annotation of bins, even where no closely related cultured genomes exists.

Present investigation

The papers included in this thesis all present, compare, or apply bioinformatic methods in metagenomics. Furthermore, the only paper not focusing on metagenomic binning is Paper I, where a reference assembly and a database for the Baltic Sea is presented. Paper II presents the software CONCOCT for metagenomic binning which is applied to a Baltic Sea time series dataset in Paper III. In Paper IV genomes reconstructed from metagenomes are compared against single-cell sequenced genomes from the same environment. Finally, in Paper V a new larger dataset from the Baltic Sea is used to reconstruct an order of magnitude more genomes compared to Paper III.

Paper I - BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea

The paper placed as the first paper in this thesis is not first in a chronological sense. Instead, it is placed first since it is not using metagenomic binning as opposed to the other papers. On the other hand, this project is the individual project I have spent most time on, as it was a core part of the BONUS BLUEPRINT project. This project was aimed at developing a framework for determining environmental conditions in marine water samples using metagenomics. One of the key deliverables of the project was a comprehensive reference metagenome for the Baltic Sea, inspired by the large-scale studies of the human microbiome. A large co-assembly accompanied by a web-based interface was constructed as a resource for other researchers, within and outside of the BONUS BLUEPRINT project. The reference assembly has been extensively used within the BONUS BLUEPRINT project to analyse additional samples, not included in the assembly construction. This is accomplished by mapping of metagenome or metatranscriptome reads to the annotated genes of the reference assembly, and thereby quickly acquire a functional or taxonomic profile of the sample.

The paper is designed to solely present data which is relevant and re-usable to the research community. Therefore, it does not contain any analysis except technical validation of the data. The main content is the reference assembly and the web-interface together with the dataset it is based on. The size of the input dataset enabled a co-assembly resulting in more than 6 million unique gene sequences. These were extensively annotated for function and taxonomy and their relative abundance in each sample was quantified. Furthermore, the publicly available web-interface which enables easy access to all this information while additionally providing search tools and some visualizations, is described.

My contributions to this paper was: I was involved in the planning and design of the project, I performed the bioinformatic analysis and implemented the database along with the web-interface. I also wrote most of the paper.

Paper II - Binning metagenomic contigs by coverage and composition

This paper presents the program CONCOCT, a method for automatic metagenomic binning, using nucleotide composition and coverage over multiple sample. The project that led up to CONCOCT originated as a master thesis project which Brynjar Smári Bjarnason and I set out to finish. This was my first acquaintance with Anders Andersson, at that time my master thesis supervisor and later the supervisor of my PhD studies. Co-supervisor of the CONCOCT project was Christopher Quince, who eventually designed the clustering algorithm and implemented most of the software. The algorithm uses a Gaussian mixture model, representing each genome with a gaussian distribution in multiple dimensions.

The performance of CONCOCT is displayed on two simulated datasets and two real datasets. Overall, CONCOCT was shown to cluster all datasets well, while a large number of genomes (N=101) were more easily handled than a smaller dataset (N=20) where closely related strains were present. When evaluating the importance of multiple samples, a general improvement in clustering performance per added sample could be observed up to around 50 samples.

Regarding my contributions to this paper, Brynjar Smári Bjarnason and I implemented the python wrapper, which includes construction of nucleotide composition vectors. Furthermore, I participated for the full duration of the project, executed parts of the comparisons and was involved in writing the manuscript.

Paper III - Metagenome-assembled genomes uncover a global brackish microbiome

Immediately when I started my PhD studies, I became involved in a project aiming at applying the CONCOCT method to a Baltic Sea time series dataset. The paper resulting from this project introduced the term Metagenome Assembled Genomes (MAGs), which has become a standard phrase in metagenomics for quality approved bins. The metagenomic binning also turned out successful, using individual-sample assemblies, 83 MAGs were reconstructed, and were de-replicated to 30 approximately species level clusters.

While the main focus of my PhD studies has been on bioinformatic methods, these are irrelevant if not applied to a biological context. Out of the five papers presented in this thesis, Paper III contains the most extensive biological interpretations. The reconstructed MAGs were shown to be most closely related to genomes found within other brackish waters even though these were geographically very distant. Furthermore, evidence showed that the adaptation to the brackish environment were in fact older than the formation of the Baltic Sea. This led to the conclusion of the existence of a global brackish microbiome.

For this paper, I am listed as the third author. My contributions were mainly within methodological aspects, such as performing and evaluating the metagenomic binning. However, I also took active part in designing the analysis and writing the manuscript.

Paper IV - Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes

In this manuscript, we set out to compare two approaches for obtaining genomes from uncultivated prokaryotes. The first method was to recover genomes from metagenomes and the second one was the more established technique of single-cell sequencing. The genomes from the first technique were the ones presented in Paper III while the single-cell sequenced genomes were obtained from the same station in the Baltic Sea, sampled approximately one year later. Genomes that were obtained from both methods turned out to be almost identical between the methods. In terms of quality, the single-cell sequenced genomes were found to be consistently less complete than the corresponding MAGs. Furthermore, the errors caused by metagenomic binning were estimated to be less than those caused by metagenomic assembly.

For this paper, where I am co-first author together with Christofer M.G. Karlsson, I was the main responsible for the bioinformatic analysis. Furthermore, I generated most figures and wrote most of the original manuscript together with Christofer.

Paper V - Recovering 2,032 Baltic Sea microbial genomes by optimized metagenomic binning

Given that the data obtained and presented in Paper I was readily available, I was hoping to find time to perform metagenomic binning using this dataset, before the end of my PhD. Luckily this was possible. It might be worth adding some more biological analysis to this manuscript. However, I think it still adds a valuable contribution to Baltic Sea microbial research with, compared to Paper III, an order of magnitude larger number of genomes and species clusters recovered.

This paper also displays how much simpler recent developments has made it to conduct a metagenomic binning study. Compared to Paper III, steps that have been simplified or improved include assembly, quantification of contigs, the actual binning program, evaluation of bins, taxonomic annotation of bins and comparison to genomes previously obtained from uncultivated prokaryotes. Furthermore, new possibilities to obtain eukaryotic microbial genomes have opened up.

For this paper, where I am listed as the first author, I have been fully responsible for processing of the raw data, performing metagenomic binning, assigning MAGs from bins, and performing taxonomic and functional annotation. Furthermore, I drafted the first version of the manuscript and have written a large part of the current version of the manuscript.

Future perspective

This thesis have presented bioinformatic methods for metagenomics, with a special focus on metagenomic binning. Metagenomic binning have had a large scientific impact. Besides drastically extending the prokaryotic tree of life it has also led to other important biological findings such as the discovery of an Archaea, seemingly related to the first eukaryote. I believe metagenomic binning still has a great potential for further discoveries, both by applying it to novel environments and by further analysis of available datasets. Possibly also, by applying methods to achieve strain-resolved binning. However, the future might also carry completely different challenges.

Current tools developed for metagenomic binning are specifically designed to fit with current sequencing technologies. The relatively short read lengths produced by current massively parallel sequencing machines limits the success of assembly and necessitates metagenomic binning to construct genomes from metagenomes. Furthermore, the massive number of reads produced by these machines is also what enables accurate quantifications of each assembled fragment. This is the foundation for current metagenomic binning methods. New sequencing techniques could therefore potentially drastically change metagenomic binning or even make it obsolete. One such promising technique is read phasing, which have been used to reconstruct genomes from metagenomes (Bishara et al. 2018). However, it remains to be seen whether phasing methods also can produce accurate estimations of relative abundance, a very important feature of metagenomic binning in order to draw ecological conclusions.

Room for possible improvements can also be seen within the bioinformatics area. One such improvement could be quantification for metagenomic binning. While Kallisto was successfully used for metagenomic quantification in Paper V, it was constructed for an entirely different task. A fast quantification method specifically constructed for metagenomics could probably be implemented in such a way that results would be more reliable. Improvements of metagenomic binning tools are probably also possible, although I believe only minor improvements in clustering performance will be achieved within the current school-of-thought. I do however, see room for improvements when it comes to the user interface offered by these tools, CONCOCT included.

A second version of the mentioned comparison of metagenomic binning tools have been announced and it promises to include more samples, which was the major issue with the original study. This could establish a new reference dataset and spark optimizations of the different tools. However, as the tools become optimized for this specific dataset, the comparative value and relevance of that dataset will likely decrease.

A more ground-breaking idea would be to include the multiple-sample abundance information directly into an assembly program. This could theoretically improve assembly drastically. In practice, however, this is very hard since assembly is already a very memory intensive task. Including more information for the program to use would need, if possible, highly competent engineering.

We are still far from a complete understanding of the microbial world. The incredible diversity of prokaryotes promises future discoveries that will potentially change the foundations of our scientific understanding. Furthermore, these discoveries could perhaps also help us deal with current challenges, such as creating a sustainable society. Whether these discoveries of the future will be mediated by metagenomic binning or entirely different techniques is, when considering this bigger picture, less important.

Acknowledgement

This would not have been possible without the support from a large number of people. First and foremost, my supervisor Anders. With your creative ideas, deep knowledge, and contagious enthusiasm you've made my working efforts much more successful. Furthermore, you've always been a great boss, which was one of the main reasons why I wanted to start my PhD studies for you. The reason I started with bioinformatics might, however, well be due to my master program lecturer, and later PhD co-supervisor, Lars Arvestad. I've always been inspired by your technical expertise and sharp mind.

I'm also very grateful for having been able to work with Christopher Quince, the main contributor to the CONCOCT project. Your brilliant mind, sense of humor and warm personality is a true inspiration.

When I started as a PhD student I became a member of a great team, the EnvGen group. Thank you Luisa, Yue, Ino, Conny, Olov, Jürg, Carlo, John, Jian, Ryno and Markus for our great collaborations, tasty fikas and a simply lovely work atmosphere. This thank you is also equally directed towards our affiliated visitors Sophie, Christin and Rene. The first year would, however, not have been as successful without Konstantin of SciLifeLab IT who managed to save my coffee-drenched brand new laptop.

To my shared first co-authors Brynjar and Christofer, I would like to apologize for the first letter of my family name, leaving you with less credit than you deserve. I also want to thank you for the fun times we had collaborating. Furthermore, all other co-authors deserve a special thanks for the time you invested in all of these different projects, to make them as good as possible.

In fact, this thesis would never had come to be without the BONUS BLUEPRINT project, who contributed with the funding for my position as well as a large share of the data used within this thesis. I would like to thank all the collaborators within this project for the great time, but especially to Lasse who managed to keep us focused at the task ahead.

I would also like to thank Daniel Lundin, who first hired me as a programmer even though I had very limited experience, and later also introduced me to Anders. You're a seemingly endless resource of knowledge and scientific enthusiasm. Someone who also hired me, for which I am very grateful, is Per Kraulis. Thank you for engaging in presentations and journal clubs, and for your overall passion for science.

Also, I'd like to thank Brynjar, Roman, Valentine, Guillermo, Mario and Ino, who were my first major programmer role models from where I found a lot of inspiration to develop as a programmer. Furthermore, the reason I did manage to get into programming is probably due to my dear friend Olof who lent me a Thinkpad x40 for several years, just to make sure I became a Linux user. Thank you!

I haven't only been working during these years. I have been very lucky in having close friends present at work and after work-hours. I'd especially like to thank the Orsa-, Halvvasa-, Helvasa-, Cycling-, Jöns Jacob- and Friskis-crews, where Anders, Michaela, Britta, Erik, Olov, Phil, Conny, Guillermo, Adria, Robin, Luisa and Yue were all major contributors. Also, since I know you will be reading this, thank you Esther for the lovely fika.

You're all too many to be mentioned individually, but I mean *thank you* to all the amazing people currently working at or have been working at NGI during these years and everybody else at SciLifeLab alfa floor 3.

Thank you also, my dear parents, grandparents, and the rest of my family: Elin, Josefin, Robin and Vera. And finally,

My beloved wife Johanna:
You are the best thing that have ever happened to me!

References

Abe, Takashi, Hideaki Sugawara, Makoto Kinouchi, Shigehiko Kanaya, and Toshimichi Ikemura. 2005. “Novel Phylogenetic Studies of Genomic Sequence Fragments Derived from Uncultured Microbe Mixtures in Environmental and Clinical Samples.” *DNA Research* 12 (5): 281–90.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. “Genome Sequences of Rare , Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes” *Nature Biotechnology* 31: 533-38.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Bishara, Alex, Eli L. Moss, Mikhail Kolmogorov, Alma Parada, Ziming Weng, Arend Sidow, Anne E. Dekas, et al. 2018. “Culture-Free Generation of Microbial Genomes from Human and Marine Microbiomes.” *bioRxiv*. biorxiv.org. https://doi.org/10.1101/263939

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, et al. 2015. “Unusual Biology across a Group Comprising More than 15% of Domain Bacteria.” *Nature* 523 (7559): 208–11.

Chatterji, Sourav, Ichitaro Yamazaki, and Zhaojun Bai. 2008. “CompostBin : A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads.” *Research in Computational Molecular Biology* Berlin, Heidelberg: Springer: 17–28.

Delmont, Tom O., Christopher Quince, Alon Shaiber, Ozcan C. Esen, Sonny T.M. Lee, Sebastian Lucker, and A. Murat Eren. 2017. “Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in the Surface Ocean.” *BioRxiv*. biorxiv.org. https://doi.org/10.1101/129791

Dick, Gregory J., Anders F. Andersson, Brett J. Baker, Sheri L. Simmons, Brian C. Thomas, a. Pepper Yelton, and Jillian F. Banfield. 2009. “Community-Wide Analysis of Microbial Genome Sequence Signatures.” *Genome Biology* 10 (8): R85.

Eddy, S. R. 1998. “Profile Hidden Markov Models.” *Bioinformatics* 14 (9): 755–63.

Falkowski, Paul. 2012. “The Power of Plankton.” *Nature* 483 (Suppl 7387): S17.

Fenchel, Tom. 2008. “The Microbial Loop – 25 Years Later.” *Journal of Experimental Marine Biology and Ecology* 366 (1): 99–103.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M.

Goodman. 1998. "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." *Chemistry & biology* 5 (10): R245-49.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2001. *The Elements of Statistical Learning*. Vol. 2. New York: Springer.

Herlemann, Daniel Pr, Matthias Labrenz, Klaus Jürgens, Stefan Bertilsson, Joanna J. Waniek, and Anders F. Andersson. 2011. “Transitions in Bacterial Communities along the 2000 Km Salinity Gradient of the Baltic Sea.” *The ISME Journal* 5 (10): 1571–79.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. “EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research* 44 (D1): D286–93.

Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14.

Hu, Yue O. O., Bengt Karlson, Sophie Charvet, and Anders F. Andersson. 2016. “Diversity of Pico- to Mesoplankton along the 2000 Km Salinity Gradient of the Baltic Sea.” *Frontiers in Microbiology* 7: 679.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11: 119.

Imelfort, Michael, Donovan Parks, Ben J. Woodcroft, Paul Dennis, Philip Hugenholtz, and Gene W. Tyson. 2014. “GroopM: An Automated Tool for the Recovery of Population Genomes from Related Metagenomes.” *PeerJ* 2: e603.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities.” *PeerJ* 3: e1165.

Kislyuk, Andrey, Srijak Bhatnagar, Jonathan Dushoff, and Joshua S. Weitz. 2009. “Unsupervised Statistical Clustering of Environmental Shotgun Sequences.” *BMC Bioinformatics* 10: 316.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Li, D., C-M Liu, R. Luo, K. Sadakane, and T-W Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76.

Lin, Hsin-Hung, and Yu-Chieh Liao. 2016. “Accurate Binning of Metagenomic Contigs via Automated Clustering Sequences Using Information of Genomic Signatures and Marker Genes.” *Scientific Reports* 6 (24175).

Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and

Dirk Gevers. 2015. “ConStrains Identifies Microbial Strains in Metagenomic Datasets.” *Nature Biotechnology* 33 (10): 1045–52.

Lu, Yang Young, Ting Chen, Jed A. Fuhrman, and Fengzhu Sun. 2017. “COCACOLA: Binning Metagenomic Contigs Using Sequence COmposition, Read CoverAge, CO-Alignment and Paired-End Read LinkAge.” *Bioinformatics* 33 (6): 791–98.

Marchesi, Julian R. 2011. “Human Distal Gut Microbiome.” *Environmental Microbiology* 13 (12): 3088–3102.

Nicholls, Samuel M., Wayne Aubrey, Arwyn Edwards, Kurt de Grave, Sharon Huws, Leander Schietgat, André Soares, et al. 2018. “Computational Haplotype Recovery and Long-Read Validation Identifies Novel Isoforms of Industrially Relevant Enzymes from Natural Microbial Communities.” *bioRxiv*. biorxiv.org. https://doi.org/10.1101/223404

Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. 2014. “Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes.” *Nature Biotechnology* 32 (8): 822-28.

Olm, Matthew R., Christopher T. Brown, Brandon Brooks, and Jillian F. Banfield. 2017. “dRep: A Tool for Fast and Accurate Genomic Comparisons That Enables Improved Genome Recovery from Metagenomes through de-Replication.” *The ISME Journal* 11 (12): 2864–68.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17: 132.

Parks, Donovan. H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Proposal for a Standardized Bacterial Taxonomy Based on Genome Phylogeny.” *bioRxiv*. biorxiv.org. https://doi.org/10.1101/256800

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, Gene W. Tyson. 2015 “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes 5.” *Genome research* 25 (7): 1043-1055.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2 (11): 1533–42.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic

Sequencing." *Nature* 464 (7285): 59–65.

Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. "DESMAN: A New Tool for de Novo Extraction of Strains from Metagenomes." *Genome Biology* 18 (1): 181.

Redin, David, Erik Borgström, Mengxiao He, Hooman Aghelpasand, Max Käller, and Afshin Ahmadian. 2017. "Droplet Barcode Sequencing for Targeted Linked-Read Haplotyping of Single DNA Molecules." *Nucleic Acids Research* 45 (13): e125.

Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, et al. 2016. "Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics." *Nature Methods* 13 (5): 435–38.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation-a Benchmark of Metagenomics Software." *Nature Methods* 14 (11): 1063–71.

Segata, Nicola. 2018. "On the Road to Strain-Resolved Comparative Metagenomics." *mSystems* 3 (2).

Segata, Nicola, Daniela Börnigen, Xochitl C. Morgan, and Curtis Huttenhower. 2013. "PhyloPhlAn Is a New Method for Improved Phylogenetic and Taxonomic Placement of Microbes." *Nature Communications* 4 (2304).

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E. Lind, Roel van Eijk, et al. 2015. "Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes." *Nature* 521 (7551): 173-179.

Stahl, D. A., D. J. Lane, G. J. Olsen, and N. R. Pace. 1985. "Characterization of a Yellowstone Hot Spring Microbial Community by 5S rRNA Sequences." *Applied and Environmental Microbiology* 49 (6): 1379–84.

Svartström, Olov, Johannes Alneberg, Nicolas Terrapon, Vincent Lombard, Ino de Bruijn, Jonas Malmsten, Ann-Marie Dalin, et al. 2017. "Ninety-Nine de Novo Assembled Genomes from the Moose (Alces Alces) Rumen Microbiome Provide New Insights into Microbial Plant Biomass Degradation." *The ISME Journal* 11: 2538–51.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes." *Genome Research* 27 (4): 626–38.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, et al. 2004. "Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment." *Nature* 428 (6978): 37–43.

Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug

Rusch, Jonathan A. Eisen, Dongying Wu, et al. 2004. “Environmental Genome Shotgun Sequencing of the Sargasso Sea.” *Science* 304 (5667): 66–74.

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson. 2016. “How to Use T-Sne Effectively.” *Distill*. https://doi.org/10.23915/distill.00002

West, Patrick T., Alexander J. Probst, Igor V. Grigoriev, Brian C. Thomas, and Jillian F. Banfield. 2018. “Genome-Reconstruction for Eukaryotes from Complex Natural Microbial Communities.” *Genome Research* 28: 569-80.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. “Prokaryotes: The Unseen Majority.” *Proceedings of the National Academy of Sciences* 95 (12): 6578–83.

Woese, Carl R., and George E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms.” *Proceedings of the National Academy of Sciences* 74 (11): 5088–90.

Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer. 2016. “MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets.” *Bioinformatics* 32 (4): 605–7.

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. “MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm.” *Microbiome* 2: 26.

Zhang, Kun, Adam C. Martiny, Nikos B. Reppas, Kerrie W. Barry, Joel Malek, Sallie W. Chisholm, and George M. Church. 2006. “Sequencing Genomes from Single Cells by Polymerase Cloning.” *Nature Biotechnology* 24 (6): 680–86.