Exploration of microbial diversity and evolution through cultivation independent phylogenomics
Our understanding of microbial evolution is largely dependent on available genomic data of diverse organisms. Yet, genome-sequencing efforts have mostly ignored the diverse uncultivable majority in favor of cultivable and sociologically relevant organisms. In this thesis, I have applied and developed cultivation independent methods to explore microbial diversity and obtain genomic data in an unbiased manner. The obtained genomes were then used to study the evolution of mitochondria, Rickettsiales and Haloarchaea.

Metagenomic binning of oceanic samples recovered draft genomes for thirteen novel Alphaproteobacteria-related lineages. Phylogenomics analyses utilizing the improved taxon sample suggested that mitochondria are not related to Rickettsiales but rather evolved from a proteobacterial lineage closely related to all sampled alphaproteobacteria.

Single-cell genomics and metagenomics of lake and oceanic samples, respectively, identified previously unobserved Rickettsiales-related lineages. They branched early relative to characterized Rickettsiales and encoded flagellar genes, a feature once thought absent in this order. Flagella are most likely an ancestral feature, and were independently lost during Rickettsiales diversification. In addition, preliminary analyses suggest that ATP/ADP translocase, the marker for energy parasitism, was acquired after the acquisition of type IV secretion systems during the emergence of the Rickettsiales.

Further exploration of the oceanic samples yielded the first draft genomes of Marine Group IV archaea, the closest known relatives of the Haloarchaea. The halophilic and generally aerobic Haloarchaea are thought to have evolved from an anaerobic methanogenic ancestor. The MG-IV genomes allowed us to study this enigmatic evolutionary transition. Preliminary ancestral reconstruction analyses suggest a gradual loss of methanogenesis and adaptation to an aerobic lifestyle, respectively.

The thesis further presents a new amplicon sequencing method that captures near full-length 16S and 23S rRNA genes of environmental prokaryotes. The method exploits PacBio's long read technology and the frequent proximity of these genes in prokaryotic genomes. Compared to traditional partial 16S amplicon sequencing, our method classifies environmental lineages that are distantly related to reference taxa more confidently.

In conclusion, this thesis provides new insights into the origins of mitochondria, Rickettsiales and Haloarchaea and illustrates the power of cultivation independent methods with respect to the study of microbial evolution.

Keywords: cultivation independent genomics, metagenomics, single-cell genomics, metagenomic binning, phylogenetics, phylogenomics, phylogenetic artefacts, comparative genomics, gene tree-species tree reconciliation, rRNA amplicon sequencing, Tara Oceans, origin of mitochondria, Alphaproteobacteria, Rickettsiales, Haloarchaea, endosymbiosis

Joran Martijn, Department of Cell and Molecular Biology, Molecular Evolution, Box 596, Uppsala University, SE-752 37 Uppsala, Sweden. Science for Life Laboratory, SciLifeLab, Box 256, Uppsala University, SE-75105 Uppsala, Sweden.

© Joran Martijn 2017

ISSN 1651-6214
urn:nbn:se:uu:diva-327310 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-327310)
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


* Equal contribution

Reprints were made with permission from the respective publishers.
Other papers by the author


Martijn J, Ettema TJG (2013) From archaean to eukaryote: the evolutionary dark ages of the eukaryotic cell; Biochemical Society Transactions (41):451-457

Viklund J, Martijn J, Ettema TJG, Andersson SGE (2013) Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade; PLoS ONE 8(11):e78858

Chen SX, Bogerd J, Schoonen NE, Martijn J, de Waal PP, Schulz RW (2013) A progestin (17α,20β-dihydroxy-4-pregnen-3-one) stimulates early stages of spermatogenesis in zebrafish; General and Comparative Endocrinology 185:1-9

Contents

Preface ................................................................................................................................. 11
   Note .................................................................................................................................. 11

Introduction .......................................................................................................................... 13

The origin of mitochondria ............................................................................................... 15

Nature of the mitochondrial endosymbiosis ..................................................................... 17

Evolution of the Rickettsiales ........................................................................................ 20

Evolution of the Haloarchaea ......................................................................................... 22

The uncultivable microbial majority ............................................................................... 24

Exploring microbial diversity ............................................................................................ 26
   Who is there? - 16S amplicon sequencing .................................................................. 26
   What are they doing? - Metagenomics ....................................................................... 28
   Metagenome assembly ................................................................................................. 28
   Who is there? - Metagenomics .................................................................................... 29
   Who is doing what? - Metagenomic binning ............................................................. 30
   Who is doing what? - Single-cell genomics ............................................................... 32

Reconstructing phylogenies .............................................................................................. 35
   A phylogenetic tree ....................................................................................................... 35
   The standard phylogenetics pipeline .......................................................................... 36
   Substitution models ....................................................................................................... 37
   Tree inference ................................................................................................................. 40
   Species tree inference ................................................................................................... 43
   Phylogenetic artefacts ................................................................................................... 44

Tara Oceans .......................................................................................................................... 46

Paper summaries .............................................................................................................. 48
   Paper I | Arcanobacter and Rickettsiaceae evolution ................................................. 48
   Paper II | The origin of mitochondria ........................................................................... 49
   Paper III | Amplicon sequencing of the rRNA operon ............................................... 50
   Paper IV | Marine Group IV archaea and the origin of Haloarchaea ... 51
Abbreviations

*Single-cell genomics and metagenomics*

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDA</td>
<td>Multiple Displacement Amplification</td>
</tr>
<tr>
<td>SAG</td>
<td>Single-cell Amplified Genome</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence Activated Cell Sorter</td>
</tr>
<tr>
<td>MAG</td>
<td>Metagenome Assembled Genome</td>
</tr>
</tbody>
</table>

*Phylogenetics*

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>GTR</td>
<td>General Time Reversible</td>
</tr>
<tr>
<td>JTT</td>
<td>Empirical rate matrix, by Jones, Taylor &amp; Thornton</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal Gene Transfer</td>
</tr>
<tr>
<td>LBA</td>
<td>Long Branch Attraction</td>
</tr>
<tr>
<td>LG</td>
<td>Empirical rate matrix, by Le &amp; Gascuel</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimate</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple Sequence Alignment</td>
</tr>
<tr>
<td>NNI</td>
<td>Nearest Neighbor Interchange</td>
</tr>
<tr>
<td>SPR</td>
<td>Subtree Pruning and Regrafting</td>
</tr>
<tr>
<td>TBR</td>
<td>Tree Bisection and Reconnection</td>
</tr>
<tr>
<td>WAG</td>
<td>Empirical rate matrix, by Whelan &amp; Goldman</td>
</tr>
</tbody>
</table>

*Taxonomic entities*

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR</td>
<td>Candidate Phyla Radiation</td>
</tr>
<tr>
<td>LECA</td>
<td>Last Eukaryotic Common Ancestor</td>
</tr>
<tr>
<td>MG-IV</td>
<td>Marine Group IV archaea</td>
</tr>
<tr>
<td>TACK</td>
<td>Thaum-, Aig-,Cren- and Korarchaeota</td>
</tr>
</tbody>
</table>

*Other*

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhAT</td>
<td>Phagocytosing Archaeon Theory</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>LCA</td>
<td>Last Common Ancestor</td>
</tr>
</tbody>
</table>
Preface

While thinking about what the content and structure of my thesis introduction should be, I was reflecting back on all the work that I have done over the past five years. How am I going to summarize all the different research topics, methods, results and conclusions in an understandable and logical manner? If there was ever an opportunity for me to properly explain what the heck I have been doing all this time, this was going to be my chance. I had no idea, however, on how exactly to achieve this. Then, when I stumbled upon a picture of my recent midsommar getaway in Fulufjället national park, it struck me. All of my thesis work follows in large lines the same workflow: from water to trees. "Geez, Joran has truly lost it now!", you might be thinking, "the PhD has taken its toll!". Perhaps you are right in thinking so. Still, I believe that it will all become clear while reading my thesis. In honor of this little eureka moment, I have decided to use the picture that summarizes my thesis work while simultaneously being artistically very appealing, as the front cover. Credits go to Giulia Pilia for taking the picture.

Note

The supplementary materials of the papers included in the thesis (published and unpublished) are available on the Google Drive directory and will be maintained until September 29:

https://drive.google.com/open?id=0B9Dzp3lOU5eeRWJuNWJYaHlvZGs
Introduction

Questions asked in this thesis are evolutionary in nature:

What is the phylogenetic origin of mitochondria?
What was the nature of the mitochondrial endosymbiosis?
How did host-association traits emerge and evolve in the Rickettsiales?
How did anaerobic methanogens evolve into aerobic Haloarchaea?

If one wants to learn more about the evolution of a certain microbial lineage, one would ideally like to step into a time machine and observe the process of evolution directly. In the absence of such a device, the next best thing is to find their modern-day close relatives and compare their traits. If done properly, it may lead to a step-by-step picture of the evolution of those traits that may closely resemble their true evolutionary history.

Following this rationale, I have identified previously uncharacterized bacteria and archaea that are closely related to the lineages in question (mitochondria, Rickettsiales and Haloarchaea) and reconstructed their genomes. Until a few years ago, this meant isolating the microbe from its natural environment, and cultivating it in the laboratory. Unfortunately, the large majority of microbes refuse to grow outside their natural habitat. Recent advances in the field of genomics circumvent the need for cultivation and allow the direct retrieval of genomic data from your microbes of interest from the environment. The genomes that were retrieved in my work are based on such methods and are derived from lakes, seas and oceans.

Genome retrieval is a costly endeavor, requiring relatively expensive deep sequencing and computationally demanding analyses. It is therefore important to carefully select those environments that harbor the microbes of interest, preferentially with relatively cheap and computationally light methods. In '16S rRNA gene surveys', environmental microbes are identified by sequencing a small part of their 16S gene. Though a powerful assay in many aspects, the captured taxonomic information is limited. In this thesis I present a new method able to capture taxonomic information of near full-length 16S and 23S rRNA genes.

To properly compare the traits encoded in the reconstructed genomes of the new microbes with those of the lineage in question, it is important to understand their underlying evolutionary relationships. These relationships are typically expressed as a phylogenetic tree, in which closely related
organisms are located on neighboring branches, while more distantly related organisms are located on far apart branches. Basing your comparison on the wrong tree can lead to wrong conclusions. Using the phylogenetic information stored within the genomes, I have attempted to reconstruct the tree that most accurately describes the relationships between the lineage in question and the newly discovered bacteria and archaea.

The tree forms the basis of the next step, in which the gene content of all considered genomes is compared, and the gene content of hypothetical ancestors are estimated. Then, by following a branch in the tree from early ancestor to recent ancestor to finally today’s organism, a picture emerges that describes the flow of genes across evolutionary history.

It is with this picture that we can return to the asked question. We can look at the hypothesized gene content of the mitochondrial ancestor and so speculate about the nature of the mitochondrial symbiosis (see "The origin of mitochondria" and "The nature of the mitochondrial endosymbiosis"). Similarly, we can infer the gene content of the ancestors of the Rickettsiales (see "Evolution of the Rickettsiales") and Haloarchaea (see "Evolution of the Haloarchaeae") and formulate hypotheses on their adaptation towards the intracellular and hypersaline environments, respectively.

In the remainder of the thesis introduction, I will expand on each of these steps in the workflow. I will first delve into the background of the stated questions, before moving on to the discovery of microbial diversity, the reconstruction of genomes, phylogenetic trees and evolutionary histories. Then I will finish with the most important findings of my thesis work with respect to the stated questions.
The origin of mitochondria

Mitochondria are organelles of eukaryotic cells that generate energy-carrying ATP molecules through the electron transport chain. From an evolutionary point of view, mitochondria appear to be an ancestral feature of eukaryotes. All characterized eukaryotes contain mitochondria or mitochondria-related organelles such as hydrogenosomes and mitosomes (Embley, 2006; Müller et al., 2012; Stairs et al., 2015), except for the oxymonad Monocercomonoides, which lost the mitochondria secondarily (Karnkowska et al., 2016).

Mitochondria exhibit a number of traits that are otherwise only found in bacteria: a circular genome that encodes bacterial genes and a bacterial-like double membrane. Furthermore, mitochondria replicate their DNA and divide independently of the eukaryotic cell. These observations prompted researchers such as Ivan Wallin (Wallin, 1926) and Lynn Margulis (Sagan, 1967) to view mitochondria as descendants of endosymbiotic bacteria. Early phylogenetic analyses performed by Carl Woese that were based on the first molecular rRNA sequence data confirmed the bacterial ancestry, and pointed towards the Alphaproteobacteria as the closest living relatives (Yang et al., 1985). Later studies that used increasingly more molecular sequence data (Lang et al., 1997; Andersson et al., 1998; Gray et al., 1999; Esser et al., 2004) confirmed the alphaproteobacterial affiliation, and also showed that mitochondria are monophyletic. This implies that all modern day mitochondria and mitochondria-related organelles have descended from the alphaproteobacterial ancestor that engaged in endosymbiosis with the progenitor of all extant eukaryotes.

Since then, researchers have attempted to link the mitochondria with a specific lineage of alphaproteobacteria. This has turned out to be a non-trivial task. Despite the increasing abundance of sequence data and steadily improving phylogenetic methods, no alphaproteobacterial lineage has been unequivocally recovered as the closest living relative to modern day mitochondria. Early reports recovered the Rickettsiales as the donor lineage (Andersson et al., 1998). These bacteria share many common features with mitochondria including their obligate intracellular life-style and reduced genomes. Though later studies also found a Rickettsiales affiliation (Fitzpatrick et al., 2006; Williams et al., 2007; Sassera et al., 2011;
Rodríguez-Ezpeleta and Embley, 2012; Wang and Wu, 2015), others have nominated the Rhodospirillales (Esser et al., 2004), the Pelagibacterales (Thrash et al., 2011), an uncharacterized oceanic lineage (Brindefalk et al., 2011) or all non-Rickettsiales alphaproteobacteria as the sister lineage (Viklund et al., 2013).

Another possibility is that the sister lineage is none of the above and has yet to be characterized. Genome sequencing efforts have prioritized cultivable and medically relevant taxa and could so have missed the sister group. This idea is not far-fetched since these "uncultivables" constitute the large majority of microbial diversity (see "Exploring microbial diversity").
Nature of the mitochondrial endosymbiosis

The mitochondrial endosymbiosis is one of the most defining events in the evolutionary history of all life. It was a key step in the evolution from 'simple' prokaryotes to 'complex' eukaryotes. Other steps include the formation of the nucleus, the Golgi, the endoplasmatic reticulum and the development of endocytosis and phagocytosis. The order in which these features have evolved is unclear, due to the lack of observed 'intermediates' (i.e. organisms that carry a subset of the complex features). The prokaryote-to-eukaryote transition has therefore been referred to as 'the evolutionary dark ages of the eukaryotic cell' (Martijn and Ettema, 2013). The mitochondrial endosymbiosis could thus have occurred at any time during this period.

Models for the endosymbiosis consist of roughly three components: the host cell, the alphaproteobacterium and the nature of their relationship (Figure 1A). Different hypotheses describing each of these components have been formulated over the years.

In the Archezoa model (Figure 1B), the host was a fully-fledged eukaryote that possessed all modern-day eukaryotic features, except mitochondria (Cavalier-Smith, 1987). The alphaproteobacterium was most likely acquired
by phagocytosis. The model was based on the existence of Archezoa, a group of eukaryotes that seemingly lacked mitochondria and formed a separate group relative to mitochondria-containing eukaryotes in early phylogenetic analyses. The model lost favor when candidate Archezoa were shown to carry organelles related to mitochondria (Bui et al., 1996; Hirt et al., 1997; Tovar et al., 1999, 2003; Williams et al., 2002), and branched within the eukaryote diversity with improved phylogenetic methods (Hirt et al., 1999; Keeling et al., 2000; Philippe, 2000).

In the hydrogen hypothesis (Figure 2C), the host was a hydrogen-dependent methanogenic archaeon, and the alphaproteobacterium was a facultative anaerobic, hydrogen producing endosymbiont (Martin and Müller, 1998). Their symbiotic relationship was metabolic in nature and based on hydrogen exchange. The model predicts that genes involved in anaerobic metabolism originated in the alphaproteobacterium and were retained in modern day anaerobic eukaryotes. However, it is unclear whether these genes were acquired from the alphaproteobacterium or have been secondarily acquired via horizontal gene transfer (HGT) from other bacteria (Hug et al., 2010; Stairs et al., 2015; Leger et al., 2016). In addition, recent phylogenetic analyses suggest that the archaeal ancestor of eukaryotes was related to the non-methanogenic TACK archaea (see below), as opposed to the methanogenic lineages of the Euryarchaeota (Cox et al., 2008; Guy and Ettema, 2011; Williams et al., 2012). Though it should be noted that the hypothesis not strictly requires a methanogenic host (Martin and Müller, 1998).

In the recently proposed phagocytosing archaeon theory (PhAT) (Figure 2D), the host was a semi complex archaeon related to TACK archaea that had evolved a primitive form of phagocytosis with which it acquired the alphaproteobacterium (Martijn and Ettema, 2013). The model was conceived after phylogenetic analyses suggested that eukaryotes shared a common ancestry with TACK archaea, rather than constituting a separate domain of life (Cox et al., 2008; Guy and Ettema, 2011; Williams et al., 2012) and the realization that TACK archaea encode a number of cytoskeleton genes that are otherwise only found in eukaryotes.

Since the formulation of the PhAT hypothesis, a number of key findings have been made. The host was most likely related to 'Asgard' archaea (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017), a sister lineage of TACK archaea. Asgard archaea represent the closest living relatives of eukaryotes in phylogenetic analyses and encode a number of membrane remodeling, vesicle biogenesis, and vesicle trafficking genes that are uniquely shared with eukaryotes. Regarding the timing of the endosymbiosis, it occurred most likely relatively late, after the host had already evolved some degree of complexity. This conclusion was drawn by a study that observed that 'stem lengths' for LECA (last eukaryotic common ancestor) genes of alphaproteobacterial origin were shorter than those for LECA genes of
archaeal origin (Pittis and Gabaldón, 2016). Yet, regarding the identity of the alphaproteobacterium, strong evidence for an affiliation with a particular group of alphaproteobacteria is still lacking (see 'Origin of mitochondria'). If the mitochondrial ancestor was related to the Rickettsiales, the symbiosis may have been pathogenic in nature. However, without solid evidence for the third component of the mitochondrial symbiosis, the nature of the relationship between the host and alphaproteobacterium remains uncertain.
Evolution of the Rickettsiales

Classically, the Rickettsiales are an order of alphaproteobacteria consisting of obligate intracellular endosymbionts of arthropods and mammals (Darby et al., 2007).

Some of them are pathogens of humans and animals. Examples include *Rickettsia prowazekii* (epidemic typhus), *Orientia tsutsugamushi* (scrub typhus), *Anaplasma phagocytophilum*, *Ehrlichia chaffeensis* and *Neorickettsia sennetsu* (ehrlichiosis), *Ehrlichia ruminantium* (heartwater) and *Anaplasma marginale* (anaplasmosis) (Renvoisé et al., 2011). They are transmitted through hematophagous hosts such as ticks, lice, fleas and mites.

*Wolbachia*, another well-characterized lineage, infects over two-thirds of all arthropods and nearly all filarial nematodes (Werren et al., 2008). *Wolbachia* are maternally transmitted through the cytoplasm of the egg cell of the host. Because of this female dependency, they have evolved various ways (cytoplasmic incompatibility, female parthenogenesis, feminization and male killing) to ensure their survival into the next generation (Werren et al., 2008).

The first Rickettsiales genomes showed the hallmark signs of reductive evolution (Andersson et al., 1998; Andersson and Kurland, 1998; Andersson and Andersson, 2001). These small (typically < 1.5 Mbp), AT-rich (< 40% GC) genomes display low coding density (< 85%) and a relatively high number of pseudogenes (Darby et al., 2007). These genomes revealed that genes related to intermediary metabolism were lost in the common ancestor of the Rickettsiales bacteria, probably because they became obsolete in the nutrient-rich environment of the host. Other genes deteriorated through genetic drift (enhanced by frequent population bottlenecks and small effective population sizes) and Müller's ratchet (activated through the lack of recombination and horizontal gene transfers). Reductive evolution does not only apply to Rickettsiales, and is found in a wide range of intracellular bacteria (Toft and Andersson, 2010). On the other hand, genes necessary for host manipulation made their entree. A *virB* type IV secretion system was most likely acquired from a gamma- or epsilonbacterium (Gillespie et al., 2010; Schulz et al., 2015), and an ATP/ADP translocator may have been acquired from a chlamydia or plastid lineage (Greub and Raoult, 2003; Linka et al., 2003; Heinz et al., 2014; Major et al., 2017).

While we have a decent idea of how intracellularity and host-dependence affected the evolution of the Rickettsiales, we have yet to understand how...
their free-living ancestor adapted to the intracellular environment. Since all Rickettsiales with a sequenced genome are strictly intracellular, we are unable to reconstruct the free-living to intracellular lifestyle transition. Genomes of intermediate lineages that branch deeply relative to Rickettsiales are necessary to gain insights into this question.

In recent years, a number of new Rickettsiales lineages have been described and sequenced. The newly proposed family Candidatus Midichloriaceae (Montagna et al., 2013) is a prime example and displays features atypical of Rickettsiales. Candidatus Midichloria survives and replicates inside mitochondria of ticks (Sassera et al., 2011), and Candidatus Jidaibacter infects an amoeba (Schulz et al., 2015). In addition, many other novel types of Rickettsiales are being discovered that do not thrive in hematophagous arthropods and mammals, but in non-hematophagous arthropods, leeches, cnidarians and green algae (Vannini et al., 2014). The classic Rickettsiales thus represent a small fraction of their natural diversity.

Obtaining genome sequences from a diverse set of Rickettsiales may provide insights into the evolutionary transition from free-living to intracellular lifestyle and the evolution of Rickettsiales in general.
Evolution of the Haloarchaea

The Halobacteria (from here on referred to as 'Haloarchaea') thrive in extremely saline environments, up to NaCl saturation. They can be found in (alkaline) hypersaline lakes, marine solar salterns, natural brines and salt rocks (Andrei et al., 2012). Examples environments include Lake Magadi (Kenya) and the Dead Sea (Horikoshi et al., 2010). They typically exhibit aerobic heterotrophic lifestyles, but many are also able to respire on alternative electron acceptors (Oren and Trüper, 1990; Oren, 1991; Bonete et al., 2008). To cope with the extreme salt concentration, Haloarchaea employ a 'salt-in' osmoprotection strategy. Potassium and chloride ions are actively accumulated from the saline environment into the cell, raising osmolarity and so preventing water from seeping out (Becker et al., 2014). Other adaptations include a highly acidified proteome and an increased genomic GC content. Acidic amino acids may prevent the formation of protein aggregates (Madern et al., 2000), while an increased GC may limit the formation of thymine dimers by UV-exposure (Dutta and Paul, 2012).

The Haloarchaea are closely related to the Methanomicrobia (group II methanogens) and are thought to have evolved from a methanogenic ancestor (Forterre et al., 2002). Unlike Haloarchaea, Methanomicrobia generally exhibit anaerobic, autotrophic lifestyles. This raises the question: what lay at the basis of this remarkable transition?

In two papers, Nelson-Sathi et al aimed to address the question. With a self-developed method, they identified over a thousand gene families in which haloarchaeal genes were placed in a monophyletic group with bacterial genes. The authors concluded that all identified gene families were acquired by the last haloarchaeal common ancestor and that the acquired genes were responsible for the transition (Nelson-Sathi et al., 2012, 2015). However, the analysis included a mere ten haloarchaeal genomes and when Becker et al added 65 genomes and re-analyzed the data, they observed a 67.2% reduction in basal acquisitions (Becker et al., 2014). Furthermore, in a response to Nelson-Sathi et al, Groussin et al convincingly laid out the weaknesses of their method, arguing that it falsely inflates the number of basal acquisitions. When re-addressing the question with an explicit evolutionary model for gene gain and loss, they found substantially fewer basal acquisitions (215), and observed a more gradual pattern of acquisitions along the Haloarchaea diversification (Groussin et al., 2016).
Yet, even with the right method, gene family gains and losses can only be inferred branch by branch. The relative timing of gene flow events within a branch are thus impossible to estimate. This is especially problematic for the branch connecting the Methanomicrobia with the Haloarchaea, because it reflects a large amount of change. What was the order of events? Did halophily predate or postdate the loss of methanogenesis? For such questions, genomic data of lineages that break the branch are necessary.

The SA1 archaea, first identified in the brine-seawater interface of the Shaban Deep in the Red Sea, are such a lineage (Eder et al., 2002). The first two members of this group were recently characterized and sequenced after being isolated from hypersaline (soda) lake sediments in Siberia (Sorokin et al., 2017). The 'Methanonatronarchaeia' exhibit aerobic, methanogenic, methylotrophic heterotrophic and halophilic lifestyles and similar to Haloarchaea, were inferred to employ a 'salt-in' osmoprotection strategy. Phylogenomic analysis confirmed the sister relationship with Haloarchaea. Methanonatronarchaea may thus represent descendants of an intermediate archaean that had characteristics of Methanomicrobia and Haloarchaea. The authors inferred that loss of classic methanogenesis genes and adaptation to hypersaline environments predated the Haloarchaea/Methanonatronarchaea divergence. Then, Methanonatronarchaea experienced heavy gene loss, while the Haloarchaea lost all methanogenesis genes and gained genes for aerobic and heterotrophic pathways (Sorokin et al., 2017).

Another intermediate lineage more closely related to Haloarchaeae are the Marine Group IV (MG-IV) archaea. MG-IV were first discovered at 3000m depth in the Antarctic Polar Front (López-García et al., 2001). Yet, genomic data is lacking and little is known about MG-IV in general. Characterization and genome sequencing of this lineage may thus be of instrumental value to understand the methanogen-to-halophile transition.
The uncultivable microbial majority

A great deal has been learned about the evolution of mitochondria, Rickettsiales and Haloarchaea in the past. However, we should be aware that our current knowledge (of microbes in general) is mostly based on just a small subset of the total microbial diversity. This subset consists of roughly two categories.

The first category consists of microorganisms that have a direct impact on society. They include pathogens of humans, livestock and crops (e.g. *Mycobacterium*, *Salmonella*, *Pseudomonas*), and those that are used for the production of various foods and drinks (e.g. *Saccharomyces*, *Lactobacillus*, *Bifidobacterium*) and for biotechnology (e.g. *Agrobacterium*). Though of significant importance, they do not represent the natural microbial diversity.

The second category consists of microorganisms that can grow under laboratory conditions. Being able to grow a microbial species in pure culture offers great benefits. It allows the study the species in great detail, without the interference of other microorganisms. We can learn about their metabolism, growth dynamics, responses to various conditions and study their shapes, sizes and cellular structure under the microscope. If a genetics system is available, we can learn about the functions of various genes. However, it has turned out that growing a microbe is extremely difficult. An overwhelming majority (estimated to be >98% of all microbes (Epstein, 2013) of microorganisms either cannot be cultivated or require a yet-to-be-discovered set of growth conditions. This phenomenon is perhaps best illustrated through a classic experiment that was done in the 1980s by Staley and Konopka. They took samples from various environments and counted their cells under the microscope, while simultaneously attempting to grow these cells on standard plates. Despite observing a rich diversity with the microscope, only few colonies were observed on the plates. The experiment result is now referred to as "The Great Plate Count Anomaly" (J T Staley and Konopka, 1985). The most likely explanation of this observation is that microorganisms seldom live in complete isolation in their natural habitat. They are rather part of complex communities of which all members rely on each other to survive. The uncultivable majority has often been called the 'microbial dark matter', in analogy to the astronomical term 'dark matter'.

In the case of the Rickettsiaceae, we can illustrate the uncultivable majority via a phylogenetic tree (Figure 2). The black branches represent all strains that have been observed through cultivation-independent techniques.
(explained in 'Exploring microbial diversity'), while the red branches represent strains that have been fully sequenced. The tree portrays the issue of the uncultivable majority quite well. While up till now one small branch has been studied in exquisite detail, there is a whole tree of branches for which very little is known. This pattern is visible across all bacteria and archaea (Rappé and Giovannoni, 2003; Wu et al., 2009; Rinke et al., 2013).

Figure 2. Maximum likelihood 16S rRNA phylogeny of all recognized Rickettsiaceae sequences in the SILVA database (December 2014). Tree includes strains for which only 16S rRNA gene sequences are available (black) and strains with genome sequences available (red)

It is clear that we are just scratching the surface of the immense microbial diversity. If we are to understand the evolutionary origins of mitochondria, Rickettsiales, Haloarchaea and other lineages of interest, we need to look beyond the cultivable and society-relevant microorganisms and explore the uncultivable majority.

In the next section, I will elaborate on recently developed methods that have made this exploration possible.
Exploring microbial diversity

To find novel microorganisms, you have to look for them in their natural environment. That could be the soil of your backyard, the bottom of the ocean, a hot spring or a riverbank. Other frequently sampled environments are the gut and skin of animals. But in principle you can look in any sample you can think of. To figure out which microorganisms live in any sample, researchers use various sequencing-based methods. I will cover these here. Each method looks at the sample from a different angle. Generally, three questions are asked:

Who is there?
What are they doing?
Who is doing what?

Who is there? - 16S amplicon sequencing

One way to answer this question may be to simply extract and sequence all DNA from a sample of interest, and compare the obtained sequences with a reference database to figure out which species are in there. This however requires that complete genome sequences are available for all the species in your sample and is unfortunately never the case.

Thus, instead of sequencing all DNA, researchers attempt to sequence a single gene from all the microorganisms present in the sample (Lane et al., 1985), and compare the obtained gene sequences with a reference database. Although the database will probably not contain the exact same sequences as those acquired from the sample, it is possible to assign a taxonomic identification to each sequence via a last common ancestor approach. For example, if a sequence has high similarity hits to various species of the genus *Escherichia* in the database, the sequence will be classified as *Escherichia*. On the other hand, if instead the sequence has lower similarity hits to a range of Firmicute species (e.g. of the genera *Lactobacillus, Clostridium, Veillonella*) the sequence will be classified as a Firmicute.

So which gene should be used? It should fulfill the following requirements: (i) be universal, i.e. present in all organisms, so no species present in the sample will be missed, (ii) be resistant to horizontal gene transfer, so no species will be mistaken for another, (iii) contain fast
evolving segments so the sequence can be classified at the genus and lower levels and (iv) contain slow-evolving segments so that the sequence can be classified at the phylum and higher levels. It turns out that the 16S rRNA gene fits these requirements. In addition, its fast evolving regions are interspersed with slowly evolving regions (Figure 3), making the gene an ideal target for PCR.

Figure 3. Schematic view of the 16S rRNA gene depicting conserved (light grey) and variable (dark grey) regions. Coordinates are based Chakravorty et al., 2007. Arrows indicate example priming locations for PCR analyses.

In a typical 16S amplicon analysis, primers are designed such that they anneal to the conserved regions that flank the variable regions (Figure 3). This allows the PCR to amplify the variable region of a large diversity of microbes present in the sample. After PCR, the amplicons are sequenced and their taxonomic affiliation is determined through comparison with a reference database. Finally, a taxonomic composition is estimated that is based on the relative proportions of reads that are affiliated with a particular taxonomy. For example, if 20% of the reads are classified as Bacteroidetes, the sample is estimated to consist of 20% Bacteroidetes. The resulting sequences can also be used in phylogenetic analyses, which can be informative on whether the sample contains microorganisms from yet unknown branches of the tree of life.

Despite its obvious power, the 16S amplicon assay has a number of limitations. The first is primer bias. Though some primer pairs are called 'universal', they miss a number of taxonomic groups. An extreme example is the recently defined 'Candidate Phyla Radiation' (CPR). Despite constituting a domain-like level of microbial diversity (Hug et al., 2016), they were never discovered by 16S amplicon surveys with standard universal primers. In addition, some taxa may be preferentially amplified over other taxa, leading to misleading taxonomic compositions. To reduce primer bias, one must accept that primers are not strictly universal, and design primers specifically for the taxonomic group of interest. This should increase the coverage for that particular group substantially. The second is polymerase error. Due to the reliance on a polymerase to amplify the 16S gene, sequences may contain erroneous base-calls that stem from polymerase mistakes. This type of error is especially impactful if the polymerase mistake occurred during an early PCR cycle. When using a conventional Taq polymerase to amplify a 1000 bp region over 30 cycles, ~68% of final amplicons will have ≥1 error
(based on ThermoFisher's PCR Fidelity Calculator). To reduce polymerase error, one can use a polymerase with proofreading capability. When using for example Phusion polymerase under the same conditions, the percentage of amplicons with ≥1 error is reduced to ~1%. The final limitation stems from PCR chimeras. During a PCR, primers are occasionally not fully extended. Incomplete strands normally anneal back to their original complementary strand in the next round, but in a small percentage of the cases they anneal to a heterologous strand. Upon extension, a chimeric fragment is formed. If chimeras are formed from two dissimilar parent sequences, they can be detected and removed from the dataset.

What are they doing? - Metagenomics

Learning about the phylogenetic identities and overall taxonomic composition of a sample is certainly useful, but it tells us nothing about what it is these taxa are actually doing. What are their metabolisms? How do they interact with each other? How do they react to changes in the environment? To answer these questions, we need information on all genes encoded by the collection of microorganisms within the sample. For example, if genes related to the oxidation of ammonia are identified in the sample, it may mean that at least one taxon in the community is able to perform ammonia oxidation.

In a standard metagenomics study, all DNA is extracted from a sample of interest and sequenced with a high-throughput sequencer. Reads are then assembled into longer contiguous sequences (contigs) and genes are predicted. Then by comparing the genes with a reference database, it is possible to infer their functions, and as a consequence the biochemical processes that occur within the community. Metagenomics thus attempts to infer the collection of biochemical processes that are performed by the community of microbes in the sample.

Metagenome assembly

Metagenome assembly is an extremely challenging problem. Depending on the complexity of the sample, it can contain up to thousands of different microorganisms, each with their own unique genome. Ideally, the assembler will reconstruct each chromosome and plasmid perfectly, each in a single contig. In reality metagenome assemblies are highly fragmented and only rarely are full chromosomes assembled in single contigs. This can be attributed to the enormous complexity of the problem. I will attempt to explain the issue with an analogy. Imagine a jigsaw puzzle with a million pieces. We need to solve it, but there are some complications: (i) we don't
know the original image (a complete genome sequence), (ii) we may not have all the pieces (parts of the genome that were not extracted or sequenced), (iii) each piece may have been altered slightly with respect to the original image (sequencing error) and (iv) identical pieces may come from different areas on the image (repeats). Now imagine collecting several hundred of such jigsaw puzzles from a game store (the sample) and throwing all pieces together in one big stack. If that was not complex enough, there are still other complications to consider: (v) some of the original images are very similar to each other (microdiversity), (vi) some images are represented by many identical copies while others are represented by few copies (relative abundance) and (vii) some identical pieces may originate from different images (conserved genes). This is usually the extent of the complexity, but some researchers go one step further and add all pieces from a few hundred similar jigsaw puzzles from the game store next door (co-assembly). It is not surprising that high-performance assemblers require up to several terabytes of memory on large computer clusters and can still crash. Still, astonishingly, they sporadically yield (near) full-length genomes in single contigs. The most important factors to a successful assembly of a particular genome from a metagenomic dataset are most likely lack of microdiversity, presence of few repeats and a relatively high abundance with complete genome coverage in the sequence data. In addition, metagenome assembly algorithms are continuously being improved (Peng et al., 2012; Nurk et al., 2017).

Who is there? - Metagenomics

Metagenomics was classically used to look at the gene content of the microbial community as a whole, and was not used for taxonomic identification. But now a new method is available that extracts the taxonomic information from the metagenome assembly. Rather than using the 16S gene, the method uses a set of fifteen conserved genes that encode ribosomal proteins. Ribosomal proteins offer the same advantages as 16S (universal distribution, high level of conservation, and scarce HGT (Sorek et al., 2007)), while additionally carrying more phylogenetic information. The additional information comes from using fifteen genes rather than one, and using protein sequence with twenty characters rather than DNA or RNA sequence with four characters. The complication is however that you have to be sure that the fifteen separate genes all stem from the same genome. It turns out that these particular genes are frequently co-located in the str-spcl cluster (M Nomura et al., 1977). It is thus highly likely that these genes are found together on a single metagenome contig. Once such 'ribocontigs' are identified, one can compare their ribosomal protein sequences with their orthologs in a set of reference taxa, and construct a phylogenetic tree.
(Castelle et al., 2015). The placement of all identified ribocontigs then gives a picture of the overall phylogenetic diversity that is present in your sample. In our research group we refer to this method as 'the RP15 pipeline', and is now a standard tool for all our metagenomics projects. For an example of an RP15 tree, see Figure 1b of Paper II.

The RP15 method is very powerful. It provides the user in one image (a phylogenetic tree) with information on who is there, and how distantly they are related to each other. One can even run RP15 on multiple metagenome assemblies at the same time and check whether or not the same lineage is present in multiple samples, and if it exhibits a degree of microdiversity. This information is very useful later on in the metagenomic binning process (see 'Who is doing what - metagenomic binning').

Compared to the 16S amplicon assay described earlier (see 'Who is there? - 16S amplicon sequencing) the RP15 assay offers one clear advantage: it does not rely on PCR. This means that primer bias, polymerase errors and chimera formation are no longer a problem. On the other hand, the RP15 assay can only detect microorganisms that are assembled and encode the str-spc cluster. Further, the assay does not provide an estimate on relative abundances, requires expensive deep sequencing and is reliant on a computationally demanding metagenome assembly.

Up till now I have described methods that attempt to answer the questions 'who is there?' and 'what are they doing?', but the million dollar question is of course: 'who is doing what?'

Who is doing what? - Metagenomic binning

In the metagenome assembly we have now identified contigs containing genes that are informative on the biochemical processes of the microbial community, and contigs containing genes that are informative on the taxonomic makeup of the community. Yet, we have no idea about which contigs are derived from the same genome, the information that tells us which microorganisms are responsible for what biochemical processes.

Figure 4. The process of retrieving metagenome assembled genomes (MAGs or bins) from metagenome primary sequence data.
It is this puzzle that metagenomic binning attempts to solve (Figure 4). Different binning methods have been developed in recent years, but they are in general based on the same principles. There are a number of properties of contigs that can be used to link them with other contigs of the same genome. The first is nucleotide composition. It turns out that prokaryotic genomes generally have the same composition across the genome (Sandberg et al., 2003; Pride et al., 2003), and that this composition is unique for each genome. Thus, if two contigs share a very similar composition, they are likely derived from the same genome. Binning tools usually express the nucleotide composition as a profile of all possible tetranucleotide frequencies. Since there are 128 possible tetramers (when considering a tetramer equivalent to its reverse complement), the frequency profile is very distinctive for a genome. The second is read coverage. Each microorganism has a particular abundance in a given sample, and this abundance is reflected in the reads. Thus contigs that, per basepair, are covered by the same number of reads are likely to be derived from the same genome. This might be tricky when two or more microorganisms have very similar abundances. In such cases it is very useful to add coverage information from other samples. Contigs from two or more different microorganisms might have similar coverage in one sample, but are likely to have a different coverage profile over two samples. The more samples are considered, the more distinctive the coverage profile becomes. The caveat is that the all the different samples must at least partially overlap in their taxonomic compositions. Studies that make use of differential coverage binning therefore use samples that are related to each other. Examples include time-series, different extraction methods and different depths. The third property is read linkage. When paired-end sequencing is used, it is possible to link contigs to each other based on read pairs. If one read of a pair maps to one contig, and the other to another contig, we can be fairly certain that both contigs are derived from the same DNA molecule. The properties discussed so far attempt to recognize contigs derived from the same genome. It is of course also useful to have information on which contigs are not derived from the same genome. That information can be gained by checking the presence of highly conserved single copy genes. These are genes that, across the tree of life, typically only occur once per genome. Examples include genes coding for ribosomal proteins, RNA polymerase subunits and translation initiation and elongation factors. Thus, if two contigs contain the same single copy gene, it is likely that they are derived from different genomes. Out of all discussed properties, this is probably the least distinctive because it is based on an assumption that may not hold for a number of new lineages.

Metagenomic binning is just a few years old, but has already proven to be very powerful. The method usually recovers a large fraction of the genome, and for many genomes. Every now and then a paper is published that claims to have reconstructed up to a few thousand new genomes (Brown et al.,
2015; Anantharaman et al., 2016; Delmont et al., 2017). Major findings made through metagenomic binning include the discovery of CPR bacteria (Hug et al., 2016) and Asgard archaea (Zaremba-Niedzwiedzka et al., 2017).

Yet, binning is not all-powerful, and faces a number of challenges. As for metagenome assembly, microdiversity is a major issue. Contigs stemming from closely related strains will exhibit extremely similar nucleotide compositions and differential read coverage. In addition, one read of a read pair may be incorrectly mapped to a contig of one strain, while the other maps to a contig of another strain, creating a false link between the two contigs. Another issue is the general lack of rRNA genes within bins (Hugenholtz et al., 2016; Nelson and Mobberley, 2017). rRNA genes are highly conserved at the nucleotide level, are difficult to assemble and display deviant nucleotide compositions compared to the rest of the genome (Galtier and Lobry, 1997; Ghurye et al., 2016). In addition, contigs with rRNA genes may have elevated coverage levels due to reads from non-assembled genomes mapping onto these contigs. Finally, despite sophisticated algorithms that take as much information as possible into account, there is a risk of contamination. Bins derived from automated tools such as CONCOCT (Alneberg et al., 2014) often contain contigs from unrelated species and need to be manually curated. If no care is taken such contaminated bins may end up in public genome databases. For this reason, it is important to check the quality of metagenomic bins from previous studies before using them in new analyses.

It is important to note that genome bins should be interpreted as 'population-level genomes', especially in the presence of microdiversity. Sequences defined by the contigs are most likely not identical to the true genomes of any of the strains, but rather represent a consensus of the nucleotide variance across the strains.

**Who is doing what? - Single-cell genomics**

Next to 16S amplicon sequencing and metagenomics, there exists a third method for exploring microbial diversity independent of cultivation: single-cell genomics. Like metagenomic binning, single-cell genomics tells us "who is doing what".

The idea is straightforward: isolate a single cell, extract its DNA and sequence the genome. It is immediately clear why this method has huge potential. The main issue of metagenomics and metagenomic binning, microdiversity, does by definition not apply. The genomic sequences obtained are not a population-level consensus, but the actual sequence of a particular cell's genome. Single-cell genomics is thus the perfect tool for studying microdiversity.
In a typical single-cell genomics study (Figure 5), cells are extracted from an environmental sample. Cells are sorted into individual wells, usually with a fluorescence-activated cell sorter (FACS), each cell is lysed and DNA is extracted. Because a single cell only contains femtograms of DNA, it is necessary to amplify the DNA before it can be sequenced (Hutchison and Venter, 2006). This is done via the multiple displacement amplification (MDA) reaction, and can increase the amount of DNA to the microgram level. Cells of interest are identified through sequencing of their 16S rRNA gene, and sequencing libraries are prepared.

**Figure 5. The standard single-cell genomics workflow**

MDA uses the Φ29 DNA polymerase, a proofreading polymerase found in the *Bacillus subtilis* phage Φ29 (Lasken, 2012). The polymerase is able to amplify DNA isothermally at 30 °C due to its unique strand-displacement activity. Random hexamers anneal to different regions of the genome and are elongated by the polymerase. Upon encountering a downstream starting site, the polymerase displaces the newly formed strand and continues elongation. The displaced strand is now a new target for hexamers, and the process starts over. This process repeats itself over and over again until the reaction stops, resulting in an enormous increase in DNA. Due to the nature of the chemistry, the DNA product forms a hyper-branched structure (Figure 5).

MDA is an easy-to-use, highly efficient method to amplify DNA. But there are some less desirable effects. The first is amplification bias. Some
genome regions are amplified more than others. The difference in coverage between the most and the least amplified regions can be extreme. While some regions may be covered a thousand times, other regions are covered once or not at all. The strength of the bias may be inversely correlated with the amount of input material (Ellegaard et al., 2013). The resulting uneven coverage presents a difficult problem for genome assemblers - though specific single-cell assemblers have been developed to address this issue (Nurk et al., 2013) -, and typically result in fragmented, partial genomes. A second undesirable effect is chimera formation. Though named identically, the underlying mechanism of chimera formation is different between PCR and MDA. Occasionally, the strand currently undergoing extension is displaced by a downstream strand (that was displaced by the polymerase) through a reaction called "branch migration". The 3' terminus of the extending strand is now free to anneal to any nearby available 5'-strand and form a chimera (Lasken and Stockwell, 2007). The last concern is MDA's sensitivity to contamination. If by sheer misfortune the contaminating DNA is amplified disproportionally more, it may dominate the eventual MDA product.

It is important to address these issues while analyzing the data. Experience shows that using an assembler that is aware of amplification bias and chimera formation (Nurk et al., 2013), and checking the assembly for chimeric and contaminating contigs yields good results.

Single-cell genomics is more hands-on and requires more steps compared to metagenomics, and each step can introduce some form of error. Genomes are usually less complete (< 60% completeness is common). But single-cell genomics is able to do the one thing metagenomics can not, and that is to study cell-to-cell genome diversity of closely related strains.
Reconstructing phylogenies

We have explored the uncultivable microbial majority via 16S amplicon sequencing, metagenomics and single-cell genomics. The next natural question is: what is their position in the tree of life? The methods that address this question are collectively called phylogenetics. There is much in-depth material on phylogenetics (e.g. the books *Molecular Evolution: A Statistical Approach* by Ziheng Yang (Yang, 2014) and *The Phylogenetic Handbook* by Marco Salemi, Philippe Lemey and Anne-Mieke Vandamme (Salemi *et al.*, 2009), and the course *Computational Molecular Evolution* by Anders Gorm Pedersen¹). Here, I will introduce the basic concepts and elaborate on the phylogenetic methods that have been relevant for my research.

![Components of a phylogenetic tree](figure6.png)

**Figure 6. Components of a phylogenetic tree**

A **phylogenetic tree** is a hypothesis. It is an idea of how a set of sequences or organisms (taxa) have evolved and diverged over time (Figure 6). Leaves

---

¹ Previously on Coursera, but currently available as a series of lectures on his YouTube channel: https://www.youtube.com/channel/UCnEi3WTOeq7iDTBkHF1LYA
represent the modern day taxa or homologous sequences, nodes represent hypothetical ancestors, and branches represent the line of descent from one node to another node or leaf. The root represents the hypothetical lineage that was the ancestor to all considered taxa. Branch lengths generally represent the estimated number of substitutions that have occurred between nodes or between a node and a leaf. Alternatively they may have no meaning (e.g. in a cladogram) or represent the amount time (e.g. in fossil-calibrated trees). Finally, the topology refers to the evolutionary relationships between the taxa, irrespective of their branch lengths.

The main assumption of this hypothesis is that closely related taxa accumulated relatively fewer substitutions in their sequences than distantly related taxa, given equal overall substitution rates.

The goal of phylogenetics is to reconstruct the tree that most accurately describes the evolutionary relationships of a set of taxa or homologous sequences, in terms of topology and branch lengths.

The standard phylogenetics pipeline

The reconstruction of a phylogenetic tree virtually always follows the same step-by-step plan:

1. Gather sequences
2. Align the sequences
3. Infer phylogenetic tree

Sequences that are related to each other are normally obtained through local alignment searches -BLAST (Altschul et al., 1990, 1997)- or hidden Markov model searches -hmmer (Eddy, 1998)-. Alternatively, one can download predefined related sequences from online databases such as EggNOG (Huerta-Cepas et al., 2016) or one can gather them manually from publications.

Once gathered, a multiple sequence alignment (MSA) is constructed. The alignment algorithm attempts to identify sites (nucleotides or amino acids) in genes or proteins from different taxa that are homologous. That is, sites that are derived from the same ancestral site. It is this information that the phylogenetic software relies upon to identify and infer the number of substitutions that occur within a site over time. The MSA is thus a hypothesis that describes which sites are homologous to each other.

From the MSA a tree can be inferred. Many different methods have been developed over the years, but here I will focus on maximum likelihood and Bayesian inference methods that use probabilistic substitution models.
Substitution models

The substitution model is arguably the most important factor in phylogenetic reconstruction. It attempts to explicitly describe the substitution process in a mathematical framework. The substitution process is highly complex, and it is practically impossible to take all factors (see below) into account. The goal of the model is therefore to identify those factors that affect the substitution process the most. The process is viewed as a long chain of 'states' in which the substitution probability from the current state to the next state only depends on the current state. Because of this Markov property, substitution models are sometimes referred to as 'Markov models'. Two types of models are covered here: nucleotide and amino acid substitution models.

Nucleotide substitution models
The simplest nucleotide substitution model, introduced by Jukes and Cantor in 1969-JC69 (Jukes and Cantor, 1969)-, specifies a single free parameter that describes the overall substitution rate. It assumes that during evolution all possible nucleotide substitutions are equally probable. This assumption is violated in most cases. For example, transitions (purine-to-purine or pyrimidine-to-pyrimidine) are more likely to occur than transversions (purine-to-pyrimidine or pyrimidine-to-purine). The K80 model (Kimura, 1980) attempts to capture this by replacing the single parameter of JC69 with two substitution parameters, one for transitions and one for transversions. The HKY85 model (Hasegawa et al., 1985) adds another layer of realism by introducing equilibrium frequencies for each nucleotide. Equilibrium frequencies specify the frequencies at which nucleotides occur in the chain of substitutions over time and were assumed to be equal by JC69 and K80 models. The general time reversible model (GTR) (Yang, 1994) relaxes the constraints of the model even further by specifying separate substitution rates for all possible substitutions. As the name states, GTR is time-reversible and specifies that the substitution rate from nucleotide \( i \) to nucleotide \( j \) is equal to the rate from nucleotide \( j \) to nucleotide \( i \). Time-reversibility is not realistic but mathematically convenient.

Amino acid substitution models
The GTR model can in principle be applied to protein alignments as well. In that context, the model specifies separate substitution rates for all possible amino acid substitutions and separate equilibrium frequencies for all amino acids. This results in a much larger number of free parameters (208) compared to the GTR model used for nucleotide data (9). To reliably estimate this large number of parameters would require a large dataset, and a large amount of computational resources. It is therefore preferred to use
empirical models (or matrices). Here the relative substitution rates (or exchangeabilities) and equilibrium frequencies are predefined and based on the comparisons of large quantities of sequence data. The first empirical model was constructed by Margaret Dayhoff and colleagues (Dayhoff et al., 1978) and was based on 1572 substitutions in 71 groups of closely related proteins. The model parameters were estimated by reconstructing ancestral sequences from very similar proteins using parsimony, and then tabulating amino acid substitutions along the inferred tree. The Dayhoff matrix was updated to the JTT model by Janet Thornton and colleagues in 1992 (Jones et al., 1992), using the same approach but with a much larger collection of protein sequences. Empirical models can also be constructed by estimating the parameters via maximum-likelihood search (see section "Tree inference") under the GTR model. Simon Whelan and Nick Goldman used such an approach in 2001 on 182 alignments of nuclear proteins to construct the WAG matrix (Whelan and Goldman, 2001). Seven years later Quang Le and Olivier Gascuel constructed their LG matrix with a similar method, but incorporating rate variability across sites in their inference and 3912 alignments from Pfam (Le and Gascuel, 2008).

Two observations can be made from these matrices. First, amino acids that are very similar in their physico-chemical properties substitute each other more frequently than do dissimilar amino acids. For example, aspartic acid and glutamic acid, the only residues with negatively charged side chains under physiological conditions, replace each other very frequently. Second, amino acids with near-identical codons replace each other frequently. For example, physico-chemical similar amino acids arginine and lysine replace each other frequently under the standard genetic code, while barely replacing each other under the mitochondrial genetic code (Yang, 2014).

Rate variation across sites
All the models stated above describe a single chain of substitutions that occur within a site over time. Nucleotides or amino acids replace each other with specified substitution rates, and occur in specified equilibrium frequencies. While inferring the phylogenetic tree, these models model the replacement process for all sites equally and assume that all sites evolve at the same rate. This is unrealistic. Some sites are under negative (purifying) selection and thus evolve relatively slow, while other sites are under neutral or positive selection and evolve relatively fast. To capture this across-site rate variation, substitution models are complemented with a parameter that multiplies the overall substitution rate with a specified factor. The model then lets different sites evolve with different multiplication factors. The rate distribution across sites is described by a gamma (Γ) distribution. By adjusting the associated alpha (α) parameter, a gamma distribution can take different shapes. For example, if α < 1 the distribution is L-shaped. Most sites are then modeled to have low substitution rates while few sites have
very high substitution rates. If $\alpha > 1$, the distribution is bell-shaped and most sites are modeled to have roughly equal substitution rates. Thus, by adding just a single free parameter ($\alpha$), the rate variation across sites can be modeled. The gamma distribution is strictly continuous, but is typically discretized into four categories of equal rate probabilities for computational efficiency. When a substitution model is complemented with a discrete gamma distribution for rate heterogeneity across sites, the suffix "+$\Gamma4$" is added, for example LG+$\Gamma4$.

**Mixture models**

Besides assuming that all sites evolve with the same rate, standard substitution models also assume that all sites evolve according to the same substitution pattern (defined by a set of relative substitution rates and equilibrium frequencies). This is not biologically realistic. Some sites, for example active sites in enzymes, may have a strong preference for positively charged amino acids, while completely ignoring most other amino acids. Other sites may have no preference at all and feature all amino acids equally. A standard substitution model does not model either case properly. Mixture models attempt to capture such across-site substitution variation by allowing different sites to evolve under different substitution patterns. More formally, a mixture model describes a number of distinct substitution matrices that share the same set of relative substitution rates, but each carry a unique profile of equilibrium frequencies. Two types of mixture models can be distinguished: those that infer the equilibrium frequencies directly from the data (e.g. CAT (Lartillot and Philippe, 2004)), and those take the frequencies from predefined, empirically obtained frequency profiles (e.g. C10-C60, (Si Quang et al., 2008)). Mixture models generally fit the data better compared to single-matrix models, and are more robust to long branch attraction artefacts (see “Phylogenetic artefacts”). Mixture models can also be combined with the exchangeabilities of an empirical model (e.g. LG+C60) and with gamma distribution for rate heterogeneity (e.g. LG+C60+$\Gamma4$).

**Summary**

Modeling the complex substitution process has come a long way since Jukes and Cantor developed their single parameter model. Currently, the most biologically realistic and computationally tractable model (CAT+GTR+$\Gamma4$) accounts for separate relative substitution rates between states and separate equilibrium frequencies per state (GTR), across-site substitution rate variation ($\Gamma4$) and across-site equilibrium frequency variation (CAT) simultaneously. Yet, two other important aspects of the substitution process are not integrated: within-site equilibrium frequency variation (non-stationarity) and within-site substitution rate variation (heterotachy). In other words, non-stationarity and heterotachy mean that equilibrium frequencies and overall substitution rates along the Markov chain may change over time,
respectively. The former has been modeled either separately (NDCH - (Foster, 2004), BP - (Blanquart and Lartillot, 2006) or in combination with CAT (CAT+BP - (Blanquart and Lartillot, 2008)). CAT+BP, though very attractive, is currently too computationally intractable for current day standard phylogenomics datasets.

Tree inference

Given an alignment and a substitution model, a tree can be inferred. Several categories of tree inference methods exist, but here I will focus on maximum likelihood and Bayesian inference.

Maximum likelihood

The maximum likelihood (ML) method aims to find the tree amongst all possible trees that has the highest likelihood. The likelihood $L(\theta)$ is defined as the probability $P$ of observing the data $D$ (the alignment), when the parameters $\theta$ (the tree, branch lengths and substitution model parameters) are given. Formally,

$$L(\theta) = P(D|\theta)$$

The process of searching for the tree that maximizes the likelihood of the data varies between phylogenetic tools, but is typically done as follows:

1. Propose a starting tree (usually neighbor-joining or maximum parsimony)
2. Optimize branch lengths and model parameters by maximizing the likelihood via numerical iterative algorithms
3. Rearrange the tree (see below) to yield a set of new trees
4. Maximize the likelihood for each new tree
5. Take the tree with the highest likelihood as the new starting tree
6. Repeat until no tree with a higher likelihood is found

All possible trees constitute the so-called tree space. This is a metaphor that equates all possible variations of trees to a landscape where, in principle, one specific tree or group of trees will have the best maximum likelihood score (i.e. will be a global optimum in the tree space), but where other, different trees may exist that form relatively good scores (local optima). Tree rearrangement algorithms are designed to further explore the tree space to find these peaks. Nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR) describe three different tree rearrangement operations able to explore tree space. They cut
one or more pieces of a tree and reassemble them into a different tree. NNI is a special case of a SPR operation, and a SPR operation in turn is a special case of a TBR operation. NNI thus returns a smaller set of new trees than SPR, which in turn returns fewer trees than TBR.

The ML tree found is the best amongst the visited trees, but is not necessarily the best among all possible trees. To increase the chances of finding the global optimum, tree rearrangements that explore a larger set of tree topologies (NNI < SPR < TBR) can be used, but come with increased computational cost (Swofford and Sullivan, 2003). Alternatively one can run the algorithm with many different initial starting trees in parallel, and pick the tree with the highest likelihood among all local optima.

The maximum likelihood algorithm also identifies, besides the ML tree with associated ML branch lengths, maximum likelihood estimates (MLEs) for all substitution model parameters. These can be useful when studying the substitution process.

Branch supports (i.e. a measure of statistical support for each branch of the tree) can be calculated via bootstrapping (Efron, 1992): alignment sites are resampled with replacement to yield a pseudo-alignment (a bootstrap). A large sample of independent bootstraps are constructed, and for each the ML tree is estimated. The bootstrap support for each branch in the original ML tree is then defined as the percentage of the bootstrap ML trees that contained that branch. Bootstrap branch support is thus a measure of robustness against site resampling.

Bayesian inference

The bayesian method does not attempt to find the tree with the highest likelihood, but rather a posterior probability distribution of trees that quantifies one's degree of belief in a phylogenetic tree. The posterior probability of a tree is obtained by updating our prior belief of the tree after observing the data. More formally, the posterior probability of tree \( P(\theta | D) \) is obtained by taking the product of the likelihood \( P(D | \theta) \) and the prior probability \( P(\theta) \), divided by the marginal probability \( P(D) \), which equates to the sum of all products \( P(D | \theta_i) \cdot P(\theta_i) \) over all possible values of \( \theta \):

\[
P(\theta | D) = \frac{P(D | \theta) \cdot P(\theta)}{\sum_{i}^{N} P(D | \theta_i) \cdot P(\theta_i)}
\]

The denominator serves as a normalizing constant, assuring that the addition of all possible \( P(\theta | D) \) will sum to 1. Unfortunately, the denominator is practically impossible to compute because parameter space is too large: the number of possible trees increases dramatically with the number of taxa, and the branch lengths and model parameters are continuous rather than discrete. This makes it impossible to compute the posterior probability distribution
directly. It is however possible to avoid the calculation of the denominator entirely and approximate the distribution via the Markov chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953; Hastings, 1970). MCMC works as follows:

1. Start with a random $\theta$ (i.e. random tree topology with random branch lengths and random model parameters)
2. Calculate $P(D|\theta) \cdot P(\theta)$. Call that $P_\theta$
3. Propose a new $\theta'$, that is close to $\theta$ in parameter space (tree is rearranged, branch lengths and model parameters are slightly adjusted)
4. Calculate $P(D|\theta') \cdot P(\theta')$. Call that $P_{\theta'}$
5. If $P_{\theta'} > P_\theta$ always accept proposition. If $P_{\theta'} < P_\theta$, accept proposition with probability $\frac{P_{\theta'}}{P_\theta}$. If accepted, set $\theta = \theta'$. If not accepted, set $\theta = \theta$.
6. Save $\theta$
7. Go to step 3
8. Continue until parameter space is sufficiently sampled

The effect of the algorithm is that points in parameter space near the optimum are sampled more often than points further away. The resulting frequency distribution of $P(D|\theta) \cdot P(\theta)$ is therefore an empirical approximation of the posterior probability distribution, which we are ultimately after.

Because MCMC prefers walking 'uphill' over walking 'downhill', there is the risk that an MCMC chain gets stuck around a local optimum. To increase the chances of finding the global optimum, multiple independent chains are run. When all chains are independently sampling the same optimum, hopefully the global optimum, the chains are said to have converged. Alternatively, one can initiate 'cold' chains and 'heated' chains. Cold chains move along parameter space normally, while heated chains move with larger $\theta \rightarrow \theta'$ steps. Heated chains thus explore a larger volume of parameter space, but are unable to accurately approximate the posterior probability distribution. To exploit the sampling power of a cold chain, and the explorative power of the heated chain, cold and heated chains are occasionally 'swapped'. The cold chain thus 'uses' the heated chain to escape local optima. This algorithm is called Metropolis-Coupled MCMC, MCMC^3, or MC^3 (Geyer, 1991).

Once the posterior probability distribution is properly approximated, a majority-rule consensus tree is constructed from the sampled trees. Branch supports (posterior probability support) represent the degree of belief for a particular clade and are acquired from the sampled trees. For example, if a particular branch occurs in 83% of the sampled trees, the branch and
associated clade will have a posterior probability support of 0.83. Approximated probability distributions of all model parameters can be acquired as well, if one is interested in the details of the substitution process.

One of the advantages of the Bayesian method compared to the maximum likelihood method is that the tree search algorithm is per iteration computationally less demanding. Branch lengths and model parameters are not optimized every iteration, but are sampled along with the tree topology. This allows the Bayesian method to implement more realistic, parameter-rich models.

Species tree inference

The methods described above are mainly used to describe two types of evolutionary histories: those of individual gene families, in which the tree is informative on ancestral duplications, horizontal gene transfers, losses and speciation events; and those of species, in which the tree is informative on phylogenetic relationships between species. To infer a species tree, one should ideally use a gene family that has evolved purely via speciation events and has not been lost by any of the considered species. Indeed, the earlier discussed 16S is such a gene (see "Who is there - 16S Amplicon sequencing"). Although phylogenies based on 16S have led to major insights into inter-species relationships in the past (Woese and Fox, 1977; Yang et al., 1985; Kumar and Rzhetsky, 1996) they often fail at confidently determining deep phylogenetic relationships. This can be attributed to stochastic sampling error caused by the relatively limited number of phylogenetically informative sites (Rodríguez-Ezpeleta et al., 2007; Philippe et al., 2011) present in 16S alignments. To reduce sampling error, it is common practice to increase the number of informative sites by concatenating alignments of separate genes into larger supermatrix alignments. This can be done either with rRNA genes whereby the 16S and 23S genes are concatenated (Williams et al., 2012; Ferla et al., 2013; Zaremba-Niedzwiedzka et al., 2017), or with protein-coding genes whereby highly conserved, purely speciating genes are combined into large multi-gene alignments (Brown et al., 2001; Wolf et al., 2001; Brochier et al., 2002). Phylogenetic methods that use genomic data to infer inter-species relationships are collectively called phylogenomics. Though concatenation reduces sampling error, it may increase systematic error and could lead to highly supported but incorrect species trees (Rodríguez-Ezpeleta et al., 2007). Such trees are said to be affected by phylogenetic artefacts. Another source of error in concatenations are genes that were acquired not by speciation, but by duplication or horizontal gene transfer. It is thus necessary to generate and inspect gene trees for each of the genes used in the concatenation (Philippe et al., 2011).
Phylogenetic artefacts

Systematic errors occur when the assumptions of the chosen substitution model are violated (Philippe et al., 2011). For example, most models assume that the same relative substitution rates apply to all branches of the tree. In reality, some branches may have experienced accelerated substitution rates while others may have experienced decelerated rates (across-taxa rate heterogeneity). The model may underestimate the number of substitutions between fast-evolving lineages and infer that they are more closely related to each other than they really are. This phenomenon is called long branch attraction (LBA) (Felsenstein, 1978). Most models also assume that the same equilibrium frequencies apply equally to all branches of the tree (stationarity), in effect expecting that all leaves exhibit the same base or amino acid compositions. Because modern day orthologous genes and proteins vary in their compositions (across-taxa compositional heterogeneity), we know that this expectation is highly unrealistic. As a result, unrelated lineages with similar compositions may falsely branch together in the tree, a phenomenon known as compositional bias (Foster et al., 1997).

To prevent artefacts one must evade model violation. This can be done by choosing a more realistic, better fitting model. If this is not possible (due to an absence of such models or computational restrictions), one must transform the data in such a way that it no longer violates the model, without disrupting the underlying phylogenetic signal. This can be done in a number of ways: (i) remove taxa associated with long branches, (ii) removal of sites that contribute most to the model violation and (iii) recoding the alignment. In addition, adding taxa that break long branches may lead to better estimate of the substitution process and thus a more accurate tree. When additional taxa are not available, and the long-branch taxa are pertinent to the phylogeny, site removal and recoding are the only available options. Recoding reduces the alphabet from 20 to 4 characters (in case of amino acid data) or from 4 to 2 characters (in case of nucleotide data) based on groups of residues that typically substitute each other. Groups can be defined based on physico-chemical properties (Dayhoff4: AGPST, DENQ, HKR & FWYILMV, C is translated to missing data - (Rodríguez-Ezepeleta et al., 2007), Dayhoff6: AGPST, DENQ, HKR, ILMV, FWY and C - (Hrdy et al., 2004), RY recoding: AG & CT - (Woese et al., 1991) or can be empirically obtained (SR4: AGNPST, CHWY, DEKQR & FILMV - (Susko and Roger, 2007). Recoding also reduces across-taxa compositional heterogeneity by replacing different amino acids that each are favored by AT-rich or GC-rich genomes with the same character (Rodríguez-Ezepeleta et al., 2007). It further allows for the detection of substitutions occurring between groups that are otherwise masked by multiple substitutions. The cost is a reduction
in overall phylogenetic signal, due to fewer informative sites and fewer inferred substitutions. It should be noted that although these transformations alleviate artefacts, they do not guarantee the correct tree. Transformed alignments may still violate the model, though to a lesser degree.

Co-localizations of fast-evolving branches and/or compositionally similar taxa in the tree are not necessarily false. To investigate whether such relationships are the result of systematic error or true phylogenetic signal, one can apply any of the data transformations described above and check for major topological changes in the resulting tree. However, observing no changes does not automatically mean the absence of systematic error. Another way to assess model violations is via posterior predictive tests. They can be performed when trees are inferred in the Bayesian framework (Bollback, 2002). Briefly, the test simulates alignments based on model parameter configurations sampled from the MCMC chain. It then calculates a particular test-statistic from each simulation, yielding a null-distribution. The test-statistic is then calculated from the original alignment and compared against the null-distribution with a one-sided test. If the observed test-statistic is considered significantly different from the null, the model is said to have inadequately modeled that property of the substitution process that the test-statistic reflects. For example, the test-statistic 'maximum square difference between global and taxon-specific empirical frequencies' (Lartillot et al., 2013), reflects across-taxon compositional heterogeneity. By comparing the posterior predictive test result of the tree inference based on the original alignment with inferences based on transformed alignments, one can assess whether the transformations have reduced model violations.

In the maximum-likelihood framework, model fits can be compared between models via hierarchical likelihood-ratio tests (hLRT) (Posada and Crandall, 1998), Akaike information criterion (AIC) (Akaike, 1974), corrected AIC (AICc) (Hurvich and Tsai, 1993), Bayesian information criterion (BIC) (Schwarz, 1978) and decision theory (DT) (Minin et al., 2003). The Bayesian framework offers in addition to posterior predictive tests also Bayes factors (Kass and Raftery, 1995) and cross-validation tests (Lartillot et al., 2007). An in-depth overview of model selection tests can be found in (Johnson and Omland, 2004; Sullivan and Joyce, 2005; Posada et al., 2004). If, according to said tests, a complex model fits the data significantly better than the simpler model, the extra parameters of the complex model reflect important facets of the substitution process for that data. In this comparison, we learn that the simpler model does not adequately explain the substitution process and thus has its assumptions violated by the data. However, it does not necessarily mean that the complex model's assumptions are not violated.
Tara Oceans

Exploration of microbial diversity is usually done by single research groups that are interested in certain type of microorganisms. They collect a number of environmental samples that they believe contain their favorite bugs and apply single-cell genomics or metagenomics or both. But there are more coordinated efforts that exist purely for the sake of discovery. One such effort is the Tara Oceans project (Bork et al., 2015; Sunagawa et al., 2015). Since 2009, volunteers have scoured the world's seas and oceans on a schooner, collecting a truly astounding amount of samples along the way (Figure 7). Many new organisms were discovered, including small ‘macro’-organisms, microorganisms and viruses. Samples were taken from depths up to 1 km, although most were taken between 5 and 200 m depths. All samples were subjected to high-throughput imaging, DNA sequencing and RNA sequencing. Over 10 Tbp of sequence data was collected, distributed over 29 billion sequences.

Figure 7. Tara Oceans sampling stations. Original figure by (Sunagawa et al., 2015)

The data was publically released in 2015 and has since been studied by many researchers across the world. While on the hunt for new alphaproteobacterial genomes I thought that the Tara Oceans data could be a useful repository. The data has turned out to be an excellent resource for a number of reasons:
(i) oceans are rich in alphaproteobacteria, (ii) most samples were sequenced deeply with Illumina HiSeq, (iii) a large number of samples shared similar microbial compositions, meaning that closely related strains could be separated through differential coverage binning, (iv) draft metagenome assemblies were available (Sunagawa et al., 2015), which meant it was possible to screen for the most promising samples with the RP15 pipeline and (v) there was a large amount of metadata available for each sample.

How to extract new alphaproteobacterial genomes from this treasure trove? The total dataset was too large to assemble, and assembling all samples individually would be too computationally demanding and too time consuming. Together with a master's student (Julian Vosseberg), we screened all draft metagenome assemblies with the RP15 pipeline (see "Who is there? - Metagenomics") and selected the most promising samples. We generated metagenome assemblies with metaSPAdes (Nurk et al., 2017) and extracted alphaproteobacterial bins. We reconstructed thirty new, high-quality alphaproteobacterial genomes distributed over fourteen novel lineages. We reconstructed another five genomes belonging to Marine Group IV archaea as well. The Tara Oceans data has been of immense value for Paper II, and has kick-started the projects described in papers IV and V. Other research groups had similar ideas and reconstructed hundreds to thousands of new genomes from the Tara Oceans data in parallel (Delmont et al., 2017; Tully et al., 2017).
Paper summaries

Paper I | *Arcanobacter* and Rickettsiaceae evolution

**Aims**
The Rickettsiaceae are well known for their obligate intracellular pathogens of humans and animals that are transmitted through hematophagous arthropods. However, other Rickettsiaceae cover a much larger diversity and infect a broader range of hosts (protists, cnidarians and leeches). Their overall evolution is poorly understood because genome sequencing efforts have heavily prioritized pathogenic lineages. When a single-cell genomics screen of a lake identified a Rickettsiaceae related cell, we had the opportunity to learn more about their evolution. Our goals were as follows: (i) sequence and reconstruct its genome, (ii) update the Rickettsiales species tree and (iii) compare its gene content with the Rickettsiaceae.

**Results**
We sequenced and assembled the single-cell amplified genome (SAG), derived from Damariscotta Lake (Main, USA). Compared to other Rickettsiales it had a fairly large genome size (~1.7 Mb), a typical AT-rich genome (32.6 %GC) and a fairly high coding density (88.4%). Phylogenomic analyses confirmed the relatedness with the Rickettsiaceae and placed the SAG firmly as their sister in the species tree. We named the SAG 'Candidatus Arcanobacter lacustris' (hereafter *Arcanobacter*). Next we compared *Arcanobacter's* gene content with other Rickettsiaceae genomes. We could only infer losses (i.e. present in *Arcanobacter* but absent in Rickettsiaceae) because any gene absent in *Arcanobacter* may be an artefact of the genome's incompleteness. Our most significant finding was an array of 18 flagellar and 9 chemotaxis genes. These genes were classically considered absent in Rickettsiales, and flagellar genes were at this point otherwise only identified in *Candidatus* Midichloria mitochondrii (Sassera et al., 2011). We found that the flagellar genes were vertically inherited, implying that they were present in the Rickettsiaceae-*Arcanobacter* last common ancestor and thus lost by the Rickettsiaceae. The origin of the chemotaxis genes was less clear and may have been originated by an independent horizontal acquisition by *Arcanobacter*. We further inferred an obligate or facultative intracellular lifestyle for *Arcanobacter*, based on a set
of host-association marker genes (an ATP/ADP translocase, a virB type IV secretion system and a set of putative effectors) and showed that it was a member of a freshwater clade that is rare in the sampled biosphere.

**Paper II | The origin of mitochondria**

**Aims**

There is no consensus on the phylogenetic origin of mitochondria. Most studies agree on an alphaproteobacterial origin, but disagree on the exact origin within Alphaproteobacteria. We aimed to address the question by increasing the taxon sampling with alphaproteobacteria residing in the oceans and applying phylogenetic methods that specifically address long branch attraction and compositional bias artefacts. Our plan was as follows: (i) identify oceanic samples harboring divergent alphaproteobacterial lineages, (ii) obtain their genomic sequences via metagenomic binning and (iii) re-evaluate the phylogenetic placement of mitochondria with increased taxon sampling.

**Results**

We screened all public metagenome assemblies of the Tara Oceans project (Sunagawa *et al.*, 2015) for novel alphaproteobacterial lineages residing in the surface of the oceans. Three samples in the Pacific Ocean were selected. Another four metagenomes from deep layers of the Atlantic Ocean from an unrelated project (Landry *et al.*, 2017) were selected as well to search for alphaproteobacteria residing in the deep ocean.

After assembling the metagenomes, we reconstructed 45 metagenome-assembled genomes (MAGs) from novel alphaproteobacterial lineages using differential coverage binning. The genomes revealed that some of the new lineages were undergoing genome streamlining (small, AT-rich genomes with short intergenic spacers) while others did not exhibit such features. This may reflect the two types of lifestyle strategies that exist in the oceans: freeliving and oligotrophic (surviving in nutrient-poor environments), and patch-associated and copiotrophic (requiring nutrient-rich environments) (Luo and Moran, 2015).

We attempted to reconstruct the Alphaproteobacteria species tree with the increased taxon sampling. By using a phylogenomics dataset consisting of highly conserved genes in conjuncture with mixture models that account for across-site substitution heterogeneity (Lartillot and Philippe, 2004; Si Quang *et al.*, 2008; Huaichun *et al.*, 2017) and methods that address phylogenetic artefacts, it was clear that the novel lineages were not closely related to the known alphaproteobacteria.
To test whether any of the new lineages were related to mitochondria, we attempted to reconstruct the Alphaproteobacteria species tree, now including mitochondria. We used the same phylogenetic inference strategy, and found that none of the new lineages branched with mitochondria. Instead, we found that mitochondria branched as a sister to either Rickettsiales ('Rickettsiales-sister') or all Alphaproteobacteria ('Alphaproteobacteria-sister'). Both hypotheses were investigated with incremental site-removal strategies, approximately unbiased tests, posterior predictive tests and alternative gene datasets. All analyses pointed towards the Alphaproteobacteria-sister being correct, and suggested that Rickettsiales-sister topology was most likely a compositional bias artefact.

**Paper III | Amplicon sequencing of the rRNA operon**

**Aims**
Standard 16S rRNA amplicon sequencing yields short partial 16S gene sequences that can be used to estimate the taxonomic composition of a sample, but are of limited use in phylogenetic analyses. We aimed to develop an amplicon sequencing protocol that captures 16S and 23S gene sequences of environmental prokaryotes by exploiting the 16S-ITS-23S locus and PacBio's long read sequencing technology. Our plan was as follows: (i) develop a PCR protocol able to generate specific accurate long amplicons from the environment, (ii) benchmark the protocol with a mock community and (iii) develop a read curation pipeline that reduces the overall error rate. We further wished to compare our method with standard 16S amplicon sequencing with respect to phylogenetic resolution, and to apply the method on four environmental samples.

**Results**
Developing an amplicon sequencing protocol targeting the 16S-ITS-23S locus encountered several complications: (i) lack of universal 23S primers in the literature meant primers had to be designed *de novo*, (ii) substantially longer amplicon length requires a long range polymerase, (iii) targeting a large locus risks co-amplifying shorter aspecific loci, (iv) PacBio sequencing requires considerably more input DNA, (v) PacBio heavily favors sequencing shorter fragments over longer fragments and (vi) PacBio exhibits a higher error rate than Illumina, even with circular consensus technology. To deal with these issues, we designed a primer targeting the 23S gene, used the proofreading Q5 polymerase to generate long and accurate amplicons, performed multiple PCRs in parallel and removed aspecific short PCR products via bead cleaning.
We prepared sequencing libraries and performed PacBio sequencing. We compared the reads derived from the mock community against the reference genomes and recovered an error rate of 2.45%. To reduce the error rate, we developed a read curation pipeline that removes highly erroneous reads by exploiting certain read properties. Briefly summarized, these are low overall read quality, local stretches with low quality base-calls, PCR chimeras and long palindromic sequences known as 'siamaeras'. Removing these reads reduced the error rate to 0.64%. Though a considerable improvment, we aimed to reduce the error rate further by exploiting the random nature of PacBio errors. We argued that a consensus constructed from reads originating of the same genome would more accurately represent the true biological sequence than most individual reads. Indeed, constructing consensus reads from sets of near-identical reads reduced the error rate to 0.18%.

We then applied the method to four diverse environmental samples and performed the first phylogenetic analyses with the obtained reads. Preliminary results indicate that using near full-length 16S and 23S sequences leads to overall statistically higher supported trees compared to using solely Illumina length 16S sequences. In addition, some environmental taxa experienced major topological shifts, perhaps reflecting an increase in phylogenetic accuracy. It thus appears that our method yields more confident phylogenetic classifications for environmental taxa. However, more analyses are necessary to verify these observations.

Paper IV | Marine Group IV archaea and the origin of Haloarchaea

Aims
Haloarchaea thrive in environments with high salt concentrations (near saturation) and mostly exhibit (facultative) aerobic, heterotrophic lifestyles. Yet, Haloarchaea are thought to have evolved from anaerobic, autotrophic methanogens. What lay at the basis of this remarkable evolutionary transition? Here we sought to address this question by (i) obtaining genomic data from the closely related Marine Group IV archaea (MG-IV), (ii) reconstruct the species tree including MG-IV, a rich diversity of Haloarchaea, and the recently characterized Methanonatronarchaeia and (iii) infer gene family gains and losses along the branches towards the Haloarchaea.

Results
During the Tara Ocean screen, we identified five samples (three in the Pacific Ocean, one in the South Atlantic Ocean, and one in the Red Sea)
with MG-IV archaea. We assembled their metagenomes, performed differential coverage binning and recovered five high-quality, near-complete MAGs. They were relatively small (~1.2 Mb) and displayed a moderate GC content (~40-42%) and a high coding density (~93-94%). Similar to other marine prokaryotes, MG-IV may thus be undergoing genome streamlining. MG-IV were further found to be globally distributed and occupied exclusively marine habitats in the sampled biosphere.

Phylogenomic analysis placed MG-IV as the closest sister to Haloarchaea, subtended by the Methanonatronarchaeia. However, an alternative topology was recovered when the data was recoded. Here, the Methanonatronarchaeia were placed as a sister to all Methanomicrobia, MG-IV and Haloarchaea. The original topology may thus represent a false attraction between Methanonatronarchaeia and Haloarchaea. Both lineages may have independently acquired similarly biased amino acid compositions as they adapted to the high salt environment.

Gene tree-aware gene family history reconstruction based on the original species tree topology suggested that the transition was accompanied by a gradual loss of methanogenesis related genes, followed by a gradual gain of aerobic respiration related genes.

It should be noted that these results are preliminary. The exact placement of the Methanonatronarchaeia needs to be more thoroughly investigated, and thus gene family history reconstruction may have to be redone with an alternative topology.

**Paper V | Early evolution of the Rickettsiales**

**Aims**
Rickettsiales comprise a group of intracellular symbionts that are shaped by reductive evolution. It is unclear how their ancestors transitioned from a free-living to a host-associated lifestyle. Upon discovering lineages closely related to Rickettsiales in the Tara Oceans data, we saw an opportunity to gain insights into this transition. We aimed to (i) reconstruct their genomes, (ii) update the Rickettsiales species tree and (iii) infer gene family gains and losses along the branch leading to the last common Rickettsiales ancestor.

**Results**
During the Tara Ocean screen for novel alphaproteobacteria we identified two samples from the Pacific and South Atlantic that contained lineages related to the Rickettsiales. We assembled their metagenomes and obtained three high-quality, near-complete MAGs via differential coverage binning. Two MAGs exhibited Rickettsiales-like features and were small (≤ 1.4 Mb)
and AT-rich (≤ 29 %GC), while the third was relatively large (~2.3 Mb) and GC-rich (~52 %GC). The former seem to be members of a clade occupying various aquatic environments, while the latter is most likely member of a clade exclusively occupying marine environments.

Phylogenomic analyses that included Alphaproteobacteria and mitochondria suggested that the MAGs constituted two lineages basal to all Rickettsiales. This placement allowed us to potentially learn about genome evolutionary dynamics that underpin the free-living to host-association transition. We updated the Rickettsiales species tree with our MAGs and other recently published genomes and applied a gene-tree aware ancestral reconstruction method that is able to map gene family gains and losses along that species tree. On the Rickettsiales stem (i.e. the set of branches connecting the last common Alphaproteobacteria ancestor to the last common Rickettsiales ancestor) we observed an early acquisition of the \textit{rvh} type IV secretion system (T4SS), followed by a large loss of metabolic genes and a late acquisition of the ATP/ADP translocase and \textit{cbb\textsubscript{3}} type cytochrome \textit{c} oxidase. These preliminary observations suggest that the latest possible transition to a host-associated lifestyle is pushed back to before the Rickettsiales MAGs divergence, and that host-association predated energy parasitism. However, more in-depth analyses are required to verify these observations.


I en annan studie har jag analyserat Rickettsiales, en grupp av intracellulära bakterier som innehåller arter som kan ge upphov till sjukdomar hos djur och människor. Dessa mikrober trivs inuti eukaryota celler och kan undvika värdens immunförsvar samtidigt som de lever på näringen som finns inuti värdcellen. I denna näringsrika och isolerade miljö


Min avhandling illustrerar hur odlings-oberoende tekniker på ett kraftfullt sätt kan förse oss med nya inblickar i hur mitokondrien, Rickettsiales och Haloarkéerna evolverade. Vidare presenterar jag en ny amplikon-sekvenseringsteknik som fångar en större mängd fylogenetisk information från vilda mikroorganismer än vad som var möjligt med tidigare tekniker.
Acknowledgements

It is the year 2011. I had just finished a research project in Utrecht as part of my master's degree studying the evolution of relaxin receptors in zebrafish. During the project I became fascinated with molecular evolution, and I decided to do my next project in this topic. While browsing some potential labs, I stumbled upon the website of this peculiar fellow, Thijs Ettema, working in the far north on the origins of eukaryotes and mitochondria. Turns out he was looking for motivated students. Boom. Just like that I was sold. I sent him an e-mail and he replied the same night. "Sure, come to Uppsala and we'll do some cool science" (paraphrasing here). A few months later I joined him and Anders, another master student who had just started working on archaeal endosymbionts of ciliates living inside cockroaches, in the molecular evolution (MolEvo) lab. Here I met Lionel, Minna, Kasia, Feifei, Kirsten, Siv, Jan, Lisa, Johan and Mayank. I didn't realize it at the time, but it was going to be the start of a long journey that would ultimately culminate into this thesis.

Thijs, I would really like to thank you for giving me, a student with no formal training in bioinformatics and evolution, an opportunity to work on these fascinating topics. You, and also Caroline, Sven and Fin, have made me feel very welcome in this foreign land, and have also supported me during hard times. I very much appreciate the freedom that I felt when doing research, and the sense of direction that I occasionally lost when hitting a wall. You have already become a superstar in the field, but I have a feeling it is only just the beginning. I feel very grateful for being part of this amazing team that you have put together over the years. You could say it is a 'powerhouse' of science ;-)

In October 2012 I rejoined the lab, but now under the disguise as a PhD student. The "EttemaLab" (or as I sometimes like to call it: "TeamThijs") had just started, and consisted for a brief moment of just Thijs and me. However, it was not long before Lionel, Jimmy, Anja, Roel and Anders joined us, along with the SiCell platform members Claudia and Anna-Maria. We were now ready to rock the field.

Lionel, I would like to take a brief moment here and thank you for your amazing co-supervision. Especially in the beginning when Unix, Perl, R, general bioinformatics and phylogenetics were still foreign concepts to me, you always took the time to guide me towards practical solutions. Even now, that you have started your own team, I still feel welcome to knock on your
door when it is necessary. It would not be unfair to say that a great deal of my practical skills, but also scientific thinking, would not have been in the shape they are now if it were not for you. **Anders**, we have shared offices since my first day in Uppsala, and we have figured out the bioinformatics way of working together. You always come up with these nifty little bash tricks and Perl scripts that made my work so much easier. Whenever my brain was stuck in mashed and convoluted thoughts, you would always listen and help me straighten them out, even if they were not related to your own projects. I think we also share a love-hate relationship with teaching Molekylärbiologi och genetik. I was lucky to have you as my office mate :) Good luck with your final thesis sprint! **Anja**, you are such a talented scientist. I have always appreciated your feedback, whether it was on my writings or posters or scientific analyses. You always manage to think beyond the expected, and to express your concerns; even it goes against everyone else's thoughts. I very much admire this about you and it has led me to, hopefully, become more critical over the years. I wish you and **Pierre** the best of luck in Texel! **Jimmy**, I don't know what kind of bet you lost, but you chose to trade the sunny and warm Hawaii for the dark and cold Sweden. But I am glad that you did. You were a great colleague and I have learned so much from you. You are now back in the US with **Karli**, and I hope you are doing well and that we'll see each other again! **Claudia** and **Anna-Maria**, thank you for being great colleagues and your feedback on my odd PCR issues, all the best in the future!

Slowly but surely, the EttemaLab grew. It was time for **Kasia, Disa, Jennah, Eva, Henning, Lina, Felix** and **Maria** to join the ranks. **Kasia**, I have always very much enjoyed our discussions on deep phylogenetic problems. I think we stimulated each other forward to resolve our respective phylogenies. **Disa**, I remember your first days in the lab, trying to capture the elusive 'Otu42' as a master student. Now you are a PhD student, studying unorthodox viruses! I would like to thank you for introducing me to the world of climbing, which is now one of my favorite pass time activities. **Jennah**, I just gazed in my glass ball and foresaw a great future for you! You are a very talented scientist, with an eagle eye for spotting issues and a detectives' instinct for solving them. Thanks for always being cheerful and bringing me cliffbars from Canada ;-) **Eva**! Que pasa cabessa?? Your arrival along with the rest of the Spanish armada (**Andrea, Alejo**) to MolEvo has made me more joyful! You are a wizard in solving technical issues, be it python, R or binning related. And now you are mastering ALE, is there anything you can't crack? **Henning**, you are the rightful pope of the protistant church. I love doing bioinformatics but it is your work that is making me jealous of field biologists. Good luck with your PhD! **Lina**, thanks for your hard work on the seemingly never-ending PCR trials. I was ready throw in the towel while you never gave up. Paper III would not have been a reality without you. Also thank you for re-sparking my interest in
swimming! All the best for your new life and medicine study in Gothenburg, say hi to Marius for me. Felix, thanks for putting order into the computational side of our lab. I know us biologists can be messy and stubborn, but we need structure! Be strict with us! Otherwise no cookies for us ;) Maria, we haven't worked much together but your feedback on my PacBio amplicon issues was very useful! I should have talked to you before the project, it might have saved some issues. Good luck with everything!

At this time we were still in our old offices in B7, which was basically refurnished lab space. It was not long before there was not enough space, and we moved to our new offices in B9. This allowed Thijs to house a new wave of EttemaLab recruits: Courtney, Laura, Jonathan, Max, Martha and Tom. Courtney, I often hear you praising other scientists, while downplaying yourself a bit. Let me take this opportunity to tell that you are an-amazing- scientist yourself and a wonderful person on top of that. To put it in your words: I want to be like you when I grow up. Thanks for being a good friend and a great colleague. Laura, I thought I knew a thing or two about phylogenetics, but then you came along and put my feet back on solid ground. I really enjoy our conversations about phylogenetic methods. Figuring out their inner workings is something I love doing, and I hope we will talk more in the future. Your feedback on Paper II has been invaluable, and could not been in its current shape without you. Thanks for being a great friend and everything else! Jonathan, I always enjoy listening to your stories during lunch breaks and fika's. I hope you are having and keep on going to have a great time here! Max, with you, the EttemaLab has gained another wizard. It's truly amazing how quickly you whip up a solid script that does exactly what I want it to do. It is safe to say that Paper's III, IV and V would not have been thesis ready in time without your help. I have not found the mitochondria's ancestral lineage, so I hereby proclaim you my successor. I have high hopes for your PhyloMagnet. Martha, you were the first who dared to ask me about my defense date. But I am glad now that you did. Thanks for assisting me with the defense process! Tom, you have just started in our lab but I can tell that you are going to be a worthy successor of Lina. I wish you a great time in the EttemaLab!

During the expansion and development of TeamThijs, we've had a number of master students doing projects in our lab. Disa and Jennah followed a similar path as me and Anders, and continued as PhD students after graduation. Other students were Guillaume, Ian, Julian, Mahwash, Dries, Anna and Ilja. Guillaume, we have had so much fun back in the day. You are now a PhD student yourself in Paris with Purificación López-Garcia. You will do a great job! I hope we will meet again :) Ian, you were the first ever student that I supervised. Together we took on a very challenging project, developing an unorthodox PCR protocol for a relatively new sequencing technology. Your hard work has laid the groundwork for what has culminated into Paper III. We have also become friends during this
time. I hope your trip to China was great and I wish you all the luck in the future! **Julian**, you were my second student and together we took on the "tsunami" of sequence data generated by the Tara Oceans expeditions. To say that you did a good job would be a vast understatement. The genomes that you recovered have been of instrumental value for Paper II, and have laid the groundwork for papers IV and V. On top of that you are going to publish your own genome announcement paper soon. Not bad! Now you have started a PhD position in Utrecht with Berend Snel and you are going to work on super interesting topics. Thanks for your hard work and good luck with everything! **Mahwash**, I was really impressed by your work on the regenerative capabilities of *Stentor*, but perhaps even more by the fact that you managed to keep up with Henning! You are going to do great things during your PhD at the EBC. **Dries**, jag tror jag måste skriva någonting svenska till dig, men jag vet inte varför! Kanske eftersom 'svensk torsdag'?? You are one of the main reasons my Swedish is not absolutely terrible. Thanks for helping me out with binning, and I wish you the best in the future! **Anna**, thanks for pushing me to join the challenge class at Campus1477 and teaching me the basics of salsa. One two three ... , five six seven ... ? Piece of cake. Soon you are going to start your own PhD at SLU, good luck with everything! **Ilja**, thanks for re-invigorating my interest in speedskating and bringing me stroopwafels. Good luck with everything in the future!

While the EttemaLab was undergoing massive expansion, the rest of MolEvo that I knew from my EBC days way back had undergone some change as well. **Dani** joined the lab and **Minna, Wei** and **Mayank**, previously master students of the lab, had started their respective PhD's with **Siv**. She later recruited **Ajith, Andrea, Christian, Gaëlle** and **Karl**. In addition, **Lisa** and **Jan** recruited **Ellie, Guilherme**, and **Alejo**, respectively. **Minna, Feifei, Mayank, Ajith, Kirsten, Siv, Lisa** and **Jan**, thank you for making me feel welcome in the MolEvo lab back in EBC, but also these days. I wish all of you the best in life. **Dani**, we had just missed each other at EBC, but we made more than up for that during our time here at BMC. I have gained a lot of respect for you as a scientist and consider you a very good friend. Now that you have graduated and become a twitter superstar, you are going to join TeamThijs soon and we are going to be colleagues. Thanks for introducing me to the Spanish community, and all the best luck in the future! **Wei**, I have come to respect you a great deal over the years. Because of you I have matured a little bit; at least I think so. I was very impressed by your solo hike of the Camino de Santiago, it must have been an amazing experience. The best of luck to you in the final stages of your PhD! **Andrea**, wait no! Andreå!! <[*Censored Spanish*]!! Hahaha, we are always having so much fun together. I really like your competitive spirit and enthusiasm when we play board games (especially Jungle Speed). Your company and also that of **Ale** always puts a smile on my face, be it with
sports (climbing, swimming), Navet lunch, social gatherings or conference travel. Thanks for being such a great friend. Good luck with your half-time and the remaining PhD! Christian, I always enjoy our conversations about sports, I hope that we can run an obstacle course race together some time! Who knows, maybe I'll qualify for the qualifiers in next years Toughest! Gaëlle, Karl, Ellie and Guilherme, thanks for being great company during Friday fika's and whenever I decided to have lunch in Navet. Alejo, we always have a great time together, whether it be training for and doing the Toughest, climbing sessions or social activities. I admire your passion for music and hope you are having a great time in Sweden. Thanks for being a good friend and good luck with your upcoming half-time! Lisa, thanks for always honoring my requests regarding TA activities. I have learned a great deal from teaching many different courses.

People that know me well know that I like to be active in my free time. Initially going to the Campus1477 gym, but later also climbing, swimming, dancing and obstacle course racing. I have done all these activities with so many wonderful people, and I will take a moment here to express the joy that I feel when sharing these experiences with them. At the gym, I loved to do these "BodyPump" (formerly "Pump it") and "Spinning" classes. Then I discovered that a certain Merima was giving a class that combined both! Merima, I will repeat the first thing I said to you: you are completely nuts, but I like it! You recently graduated and are at point of writing doing an amazing trip with Fred in South America. I wish both of you all the best! Some other fools who joined me in this masochistic ritual are Laura, Ana Rosa and Ana Mingot. Ana Rosa, you are always in such a good mood and I can't help but feel uplifted in your presence. We always have a lot of fun! Thanks for giving me a few haircuts and being a great friend. Good luck in the final stages of your PhD! Ana Mingot, I always very much enjoy your company. You have a lovely home! I hope your knee is doing better these days and that we'll see each other in the gym again soon :) Another class I recently had gathered enough courage for is "Challenge". Anna, Mercè and Tiscar, what are we doing to ourselves!? I'm dead after one class and need at least a week to recuperate. But you guys go sometimes five times a week!! You are amazing. Mercè, I am always impressed by the inexhaustible amount of energy you seem to posses and your joyful spirit. I hope to join you with Challenge again in the fall. Good luck with your PhD! Tiscar, I enjoy talking to you very much and always feel uplifted by your good mood. Let's go to Allis soon! All the best with your PhD :)

Then, a few years back I discovered that Disa, a master student at the time, was a climbing instructor at the gym. I had just done some indoor rope climbing with my Dutch friends Maarten, Niels, Mark, Martijn and Lianne in the Netherlands, and I wanted to try it out in Uppsala. It was great! I soon discovered Klättercentret and started bouldering there. It was not before long the climbing craze jumped over to Alejo, Gianni, Sophia,
Ilja, Andrea and Giulia. Our climbing Wednesdays are often my highlight of the week :) Gianni and Sophia, next to being great climbing partners you are wonderful neighbors. Thanks for driving me home from Allis during winter and helping me out when I lost my keys and locked myself out. All the best with everything in the future! Giulia, I consider you one of my best friends. We share our joy for climbing and dancing lindy and for our many common friends, and we entered the stressful final stages of our PhD's around the same time. Yet, I have no doubt that you will blast everyone away with your thesis, the same way you impress me with your climbing and dancing skills. Thanks for organizing the trips to Gothenburg and Fulufjället, they have been among the highlights of my stay in Sweden. Also thanks for being my grandma in times of need ;-) Good luck with everything!

As a child I liked to swim, and I was pretty good at it at the time. I had forgotten all about how fun it was and how good it made me feel, until I joined Lina and Marius for a swim. After that I also joined Andrea and Ale a couple of times and occasionally ran into Christian. Ale, thanks for teaching me proper technique, thanks to you I can proudly say that I'm faster than the slowest Olympic athlete!

Lina, I hear that you were the one who suggested us to participate in Tough Viking. Infinite thanks!! You have started an awesome tradition. To my deep regret, I did not participate in Tough Viking, but I did in two Toughest and one Gropen Extreme. Obstacle course racing is so. much. FUN!! Alejo, Andrea, Ilja, Laura, Mercè, we are the toughest!! I should not forget the awesome time I had at Gropen Extreme (not "extreme groping", Danil) with self pro-claimed superwo... *ahum* superhorse Jorunn! Let's do it again next year! Also thanks to Christian, Gianni and Sophia, for helping us train at the legendary Östervalla training facility.

Finally, I also started doing something I thought I would never do: learning a dance. Thank you Maria and Giulia for shoving me on the dance floor during that little kulturfestivalen workshop. It made me realize how much I missed doing something music related. Maria, I'm really impressed by your exceptional kindness, cheerful mood, hard working mentality and of course your dancing skills. You inspire me to be more kind myself. Your thesis will be awesome, all the best! Through dancing I have met Ilektra, Davide, Käthe, Lisa, Jorunn, Solstickan and Joakim. Ilektra, thank you for being an amazing dance-partner. Together we learned lindy with lightning speed! Thanks for helping me find cheap clothes, being a good conversation partner and being a great friend overall. All the best for the future! Davide, you are the undisputed, uncontested, defending world champion of minigolf in Stockholm. I really enjoy your company and your humor a lot. I've never seen someone so impressed by my 'monte carlo' card shuffling skills, haha! Thanks for hosting Valborg breakfast and all the best in the future! Käthe, I am very impressed by how quickly you picked up lindy hop. Even though you are several classes behind, you are already better
than most of us! Also with bouldering, my mouth dropped to the floor by how fast you progressed in a single session. I always enjoy our conversations very much and I wish you all the best for the future. Lisa, thank you for cheerful mood, I can't help but smile in your presence. Thanks for practicing many lindy moves with me! Good luck with everything! Jorunn, we have similar names, live in the same street, go to the same gym, go to the same dance class, go to the same bouldergym and now we are both Gropen Extreme veterans! Could we bé any more different? Good luck with your studies and all activities! Solstickan and Joakim, thank you for helping me out whenever I was stuck with a lindy move and being good company in general :) All the luck for the future!

Writing the thesis is an intense endeavor. I would like to give some special thanks to a bunch of wonderful people who supported me during this time. First of all, thank you Anders for all your efforts on papers III and IV, I hope I can return the same effort to you in a few months. Then, thanks to upcoming scientific superstar slash twitter celebrity Dani for guiding me throughout this process. And to aspiring grandma slash bouldering creativity award winner Giulia, for baking me cake and giving me mental support. I would also like to thank Courtney, Lionel, Kasia, Dani, Eva and Laura for proofreading my kappa at various stages. Further thanks to Alejo and Andrea who have helped me organize the dissertation party while they have their own half-times to worry about it. Seriously guys I would have been completely overwhelmed if it were not for you. Then I would like to thank Jennah and Max for helping out with entering references and Henning and Disa for translating the Swedish summary.

I would further like to give some special thanks to Mila Mahshid. Thank you for being a close friend despite living on the opposite side of the world and giving me mental support while I was writing. Also thank you Sandra, for sending me lovely pictures and cards from your travels and being a good friend :)

I would also like to give some special thanks to my best friends in the Netherlands, Mark, Mark/Graf, Maarten/Murk, Maarten, Martijn, Lianne, Niels and Robin. I always feel instantly like the 'olden' times whenever I hang out with you guys when I'm back home. We have so much fun, it is like I never left. We have been through some very tough times together, but I have the feeling we will have LAN parties and play pudge wars in a distant future's old peoples home, when we will be 'den ouden' for realz :) To all of you, all the best and luck and happiness in the future!

Finally I want to give the most special thanks to my family: my parents Ariëtte and Pieter and my brother Luciën for not only supporting me through my PhD, but also throughout all of my life. Despite that the wind was not always in our backs, you have always made it possible for me to carry out my studies freely, be it in Zwijndrecht, Rotterdam, Utrecht or now in Uppsala. Thank you for creating an environment in which I was free to
choose my interests and supporting me in any kind of endeavor I wished to take on. Also thank you for teaching me the importance of checking things thoroughly and thinking openly yet critically; they are probably among the most important aspects of doing science. Thank you for everything.

Joran out.

*Mic drop*
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1539

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-327310