UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1083

Phylogenomics of Oceanic Bacteria

JOHAN VIKLUND

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013

ISSN 1651-6214
ISBN 978-91-554-8767-6
urn:nbn:se:uu:diva-208441

Dissertation presented at Uppsala University to be publicly examined in BMC, B41, Husargatan 8, Uppsala, Thursday, November 14, 2013 at 10:20 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Viklund, J. 2013. Phylogenomics of Oceanic Bacteria. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1083. 33 pp. Uppsala. ISBN 978-91-554-8767-6.

The focus of this thesis has been the phylogenomics and evolution of the Alphaproteobacteria. This is a very diverse group which encompasses bacteria from intraceullar parasites, such as the Rickettsiales, to freeliving bacteria such as the most abundant bacteria on earth, the SAR11. The genome sizes of the Alphaproteobacteria range between 1 Mb and 10 Mb. This group is also connected to the origin of the mitochondria.

Several studies have placed the SAR11 clade together with the Rickettsiales and mitochondria. Here I have shown that this placement is an artifact of compositional heterogeneity. When choosing genes or sites less affected by heterogeneity we find that the SAR11-clade instead groups with free-living alphaproteobacteria. Gene-content analysis showed that SAR11 was missing several genes for recombination and DNA-repair. The relationships within the SAR11-clade has also been examined and questioned. Specifically, we found no support for placing the taxon referred to as HIMB59 within the SAR11. Ocean metagenomes have been investigated to determine whether the SAR11-clade is a potential relative of the mitochondria. No such relationship was found.

Further I have shown how important it is to take the phylogenetic relationships into account when doing statistical analyzes of genomes.

The evolution of LD12, the freshwater representative of SAR11, was investigated. Phylogenies and synonymous substitution frequencies showed the presence of three distinct subclades within LD12. The recombination to mutation rate was found to be extremely low. This is remarkable in light of the very high rate in the oceanic SAR11. This is may be due to adaptation to a more specialized niche.

Finally we have compared structure-based and sequence-based methods for orthology prediction. A high fraction of the orfan proteins were predicted to code for intrinsically disordered proteins.

Many phylogenetic methods are sensitive to heterogeneity and this needs to be taken into account when doing phylogenies. There have been at least three independent genome reductions in the Alphaproteobacteria. The frequency of recombination differ greatly between freshwater and oceanic SAR11. Forces affecting the size of bacterial genomes and mechanisms of evolutionary change depend on the environmental context.

Keywords: phylogenetics, SAR11, mitochondria

Johan Viklund, Uppsala University, Department of Cell and Molecular Biology, Molecular Evolution, Box 596, SE-752 37 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-8767-6
urn:nbn:se:uu:diva-208441 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-208441)

Till mina barn, nästa helg åker vi skridskor.

List of papers

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

I Brindefalk, B., **Viklund, J.**, Larsson, D., Thollesson, M., and Andersson, S.G.E. (2007) Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. *Mol. Biol. Evol.*, 24(3):743–756.

II Brindefalk, B., Ettema, T.J., **Viklund, J.**, Thollesson, M., Andersson, S.G.E. (2011) A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. *PLoS ONE*, 6(9):e24457.

III **Viklund, J.**, Ettema, T.J., Andersson, S.G.E. (2012) Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* 29(2):599–615.

IV Zaremba-Niedzwaska, K., **Viklund, J.**, Zhao, W., Ast, J., Sczyrba, A., Woyke, T., McMahon, K., Bertilsson, S., Stepanauskas, R., Andersson, S.G.E. (2013) Single cell genomics reveals low recombination frequencies in freshwater bacteria of the SAR11 clade. *Submitted*

V **Viklund, J.**, Martijn, J., Ettema, T.J., Andersson, S.G.E. (2013) Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PLoS ONE. In press*

VI **Viklund J.**, S.G.E. Andersson. (2013) On the importance of taxonomically representative groups for the inference of adaptive traits in surface oceanic bacteria. *Manuscript*

VII Harish, A., **Viklund, J.**, Andersson, S.G.E. (2013) New protein functions evolve by expansion of ancestral fold architectures and evolution of disordered proteins *de novo*. *Manuscript*

Reprints were made with permission from the publishers.

Contents

1. Introduction

The main focus of this thesis has been the evolutionary study of a group of bacteria called the Alphaproteobacteria. The Alphaproteobacteria are interesting for a number of reasons. But I think that the major reason for studying the evolution of this group is that it is important for our understanding of the development of the Eukaryotic cell. All life that can be seen with the naked eye is eukaryotic, including us. The mitochondrion, the power-house of the eukaryotic cell[28], is thought to have originated from the engulfment of an Alphaproteobacterial cell into the cytoplasm of an archaea[27, 9]. When trying to figure out what this proto-mitochondrion was like, we need to study the evolution of the Alphaproteobacteria and try to place the mitochondria in relation to this.

Another reason for studying the Alphaproteobacteria is their versatility. Most major bacterial life-styles are represented. There are parasitic, intracellular bacteria, the Rickettsiales is a good example of these. They live inside the cells of their host and cause diseases in humans, for example typhus[2]. Other members of this group live inside insects where they manipulate the reproductive system of their hosts[30]. Several members of the order Rhizobiales are important for agriculture. They have a symbiotic relationship with legumes in which they provide nitrogen for the host[12]. Furthermore a member of this order, *Agrobacterium tumefaciens*, is used in biotechnology for transforming the genome of host plants[39]. The Alphaproteobacteria represented by the SAR11 clade is probably the most abundant group on the planet. Up to 50% of the cells in the upper ocean layers come from this small group[22]. All this makes the Alphaproteobacteria a good model system for studying genome-evolution and adaptation to different niches.

1.1 SAR11

The SAR11 was first identified by Giovannoni et al. in 1990[15]. They performed 16S sequencing from a water sample from the Sargasso Sea and identified a number of novel bacterial clades.

The first representative of the SAR11 to be sequenced was *Candidatus* Pelagibacter ubique substrain HTCC1002[14]. Surprisingly for a species that is so abundant, *Ca.* P. ubique grows very slowly, the generation time is around 2 days. Here it was found that *Ca.* P. ubique has a very small genome for a free-living organism to go together with its very small cell-size. Not only is

the genome small, it is also very compact, with an intergenic distance of only 3 base-pairs on average. It also has a number of genomic islands or hypervariable regions, named HVR1 through HVR4. It has been suggestion that the small genome of *Ca.* P. ubique is an effect of selection to minimize the cost of replication, this is called the streamlining hypothesis.

There are several more members of the SAR11-clade that has been sequenced, it is difficult to give an exact number, but when checking the NCBI website I could find genome sequences from at least 10 isolates. Since the exact taxonomic status of the SAR11 clade is not fully resolved yet it can be difficult to identify them.

SAR11 has also been found in fresh-water lakes, this subgroup of SAR11 is commonly referred to as the LD12. These are very common, but not to the same extreme extent that the oceanic ones can be found in oceans[41].

1.2 Evolutionary Inferences

> "Nothing in Biology Makes Sense Except in the Light of Evolution"
> Theodosius Dobzhansky

Below I will introduce the concepts necessary to understand the phylogenetic methods used in this thesis.

1.2.1 Homology

The concept of homology underlies all phylogenetic inferences.

To find out how species are related we need to identify characters that share a common origin, these characters are called homologous. This is contrasted with characters that are analogous, which look similar but have different origins. The canonical example for this is the wings of bats and birds, they are analogous as wings but homologous as forelimbs.

When we want to use characters to infer species evolution it is also important that the characters we use have the same evolutionary history as the species. Characters for which this is the case are called orthologs. The definition of orthologs is that it is characters that differentiated from a speciation event. An alternative to orthologs is paralogs. These are homologous characters that we can trace back to a duplication event. This is most common for genes. For example, if a gene is duplicated in an ancestral genome, these two copies are then paralogous to each other. While each copy of the duplication will be orthologous. It can be very hard or sometimes even impossible to sort this out. Another source of spurious orthology is horizontal gene transfer, first discovered in 1951[11]. Here one or several gene(s) have been transferred between two different organisms.

A common strategy for identifying orthologs is bidirectional best blast hits. Here all genes or proteins of two genomes (A and B) is blasted against each other. If protein P_A from genome A has as its best blast hit protein P_B from genome B, and that this relation also holds in reverse, that is, P_B has P_A as its best blast hit, the two proteins are thought to be orthologous. This orthology standard has been expanded so that more than two genomes can be used, one of the more famous such sets is the NCBI Clusters of Orthologous Groups (COG) database[31].

More advanced methods for inferring homology has been implemented. In this thesis we have used methods that are based on the Markov Clustering (MCL) principle[10]. Specifically as implemented by tribeMCL and its refinement orthoMCL [21]. The MCL software needs as a starting point a matrix of pairwise similarities of all proteins in the dataset. This matrix is usually based on bitscores or the logarithm of evalues from a blast run where all proteins in the dataset that is under investigation is blasted against itself. OrthoMCL is a preprocessing step to generate the input matrix to tribeMCL. It has been shown to be perform very well on benchmarks for orthology inferences[5]. The output of running these MCL programs is a clustering of all the proteins in the original dataset.

Another approach is to use a known database of orthologous groups and then try to assign proteins to these categories. One of the biggest advantages of this is that these databases in general are curated and functionally annotated. Here blast against the COG-database is not unusual. A more robust approach is to use Hidden Markov Models (HMM) instead of blast to infer the orthologous group, this is what is done when searching the pfam-database for example[3].

A further such supervised database of homologs is the Superfamily database[16]. This is a database of HMMs built from proteins classified to be on the superfamily level in the SCOP database. Proteins that belong to the same superfamily have the same three dimensional structure and are therefore evolutionary related. Since the starting point is the structure rather than the sequence it is possible to detect homologies that are undetectable by blast or other sequence-based methods.

The main drawback of using supervised methods such as pfam or superfamily to infer orthology is that the ortholog has to exist in the database for it to be detected. Though we can be more confident in the orthologs that are detected in this way.

1.2.2 Alignment

Once one or several orthologs have been detected it is time to create a sequence alignment.

All sequence-based phylogenetic analyzes start with an alignment. It is important that the alignment method manages to align homologous characters against each other. This is a very difficult problem to solve and new alignment-methods is proposed and implemented quite often. Regardless of software, all alignments *must* be visually inspected, and sometimes manually edited. It is also common to use different programs that remove regions of an alignment that is hard to align or generally unconserved.

1.2.3 Evolutionary model

As the last step before any phylogenies can be built we need an evolutionary model to describe the mutational rates of our characters. For proteins the most common models are pre-calculated matrices, but since different matrices are calculated for different set of proteins we need to select the correct one for the protein under study. Additionally, different sites in a gene often evolve at different rates, this is most often modeled with a discretized Γ-distribution.

1.2.4 Phylogenies

There are a few different methods for inferring a tree from a set of genes. Roughly they can be classified into 4 main categories, distance based, parsimony, maximum likelihood and bayesian.

In this thesis I have used maximum likelihood, mainly as implemented by RAxML[29], and a bayesian approach as implemented by phylobayes[19].

Maximum likelihood

Maximum likelihood tries to find the tree that makes our data most probable, i.e. we maximize the likelihood function $P(Data|Tree)$. This is roughly done by optimizing the branch-lengths for a tree and then calculating the likelihood. The tree is then changed in some way and a new calculation of the likelihood is done. This proceeds until no improvements to the likelihood can be found. RAxML has a number of heuristics to speed up the tree search. Among this is the what they call the CAT model, which can be seen as an approximation of the Γ-model for rate-heterogeneity. RAxML uses the CAT-model during the initial stages of tree-search but switches to the Γ-model in the end since the CAT-based likelihoods are unstable.

When constructing a phylogeny we also want to find out how well supported it is by the data. A common way of doing this is bootstrapping. This is done by resampling the original alignment a number of times (100 times or more), and then constructing trees of these pseudo-samples. Usually we want around 80% of the bootstrapped trees to contain a clade to be confident in that it is not the effect of noise.

Bayesian Inference

In a bayesian framework we assign prior probabilities, often referred to as belief, to different hypothesis, and then as new data come in we can update these probabilities. This is calculated using Bayes' theorem:

$$P(H|Data) = \frac{P(H)P(Data|H)}{P(Data)}$$

In this equation the left hand side is called the *posterior probability*, $P(H)$ is the *prior probability* and $P(Data|H)$ is the likelihood. In phylogenetics the hypothesis is of course the tree. But other parameters is also included in the hypothesis, such as the rate-heterogeneity, the distribution of branch-lengths and so forth. And the data is the alignment. In phylogenetics most priors are uninformative, that is we do not commit to any view of what the tree should look like or what the Γ-distribution should look like for example. The actual tree search is done using the Markov Chain-Monte Carlo method and one step works approximately like this. All parameters are perturbed from their current value, and then the posterior is calculated in light of these new values. If a change in the parameters increases the probability, the new values are accepted as the starting point for the next round, if the posterior is lower the probability that the new values are accepted is proportional to the relative probability decrease. Ideally, the tree-search is aborted once all of the values have reached and maintained a steady state where they only vary very little around a fixed mean with no trends in any direction. All of the trees that has been collected during the final stationary phase are then summarized into one tree. One advantage here is that we get the support values for our tree directly from the analysis, we just look at the sampled trees. The support values for a bayesian analysis is called posterior probabilities and they should ideally be above 0.95.

Phylobayes implements an evolutionary model that is different from the normal protein matrix based ones. Unfortunately this is also called the CAT model, and should not be confused with the CAT model implemented by RAxML. The Phylobayes CAT model is a method for having different protein rate-matrices for different sites[20]. This could sound similar to what the Γ-model does, but it's not. With the Γ-model we take into account that different sites could evolve at different rates. While the CAT-model tries to take into account that amino acids at different sites in an alignment have different probabilities for evolving into other amino acids. A leucine at one position might be very prone to evolve into an iso-leucine while at a different position it might be more prone to evolve into a methionine. On top of this phylobayes also uses the Γ-model of rate heterogeneity. Obviously this is a quite complex phylogenetic model so it will quite often take a long time for it to reach stationary face, in my analyzes this has taken several weeks.

A further development of the phylobayes CAT-model is the CATbp (short for break point) model[4]. This can use different rate-matrices for the same

site but at different branches in the tree. I find this to be a very interesting model, alas, it takes prohibatively long time for it to reach stationary phase.

1.2.5 Character mapping

Once the phylogeny is inferred it is possible to put the evolution of different features of the species in an evolutionary context. I have focused on genome evolution with regards to how orthologous proteins are gained, retained and lost during the course of history. Ideally what one would like to do is to infer trees of every protein family and then reconcile the gene tree with the species tree, so far this has been too computationally demanding, but that is starting to change. Instead what is done is to try to model genome evolution by simpler means. This can be done by encoding the occurrence of protein families as characters. And then using these characters together with a evolutionary model of character evolution to infer the ancestral occurrences of the families. We decided to use generalized parsimony for this. With this method costs are associated with loss, gain and possibly duplication and then the ancestral characters are assigned such that the total cost over the entire tree is as small as possible.

1.3 Metagenomics and single cell genomics

Traditional methods for investigation of bacteria and other microbial life have been dependent on our ability to grow them in the lab. In general it is estimated that only 1% of the microbial life is culturable in the lab[1]. Thus most of life on earth is not accessible by traditional means.

There are two main techniques that are independent of culturing that has been developed to remedy this.

1.3.1 Metagenomics

Metagenomics[18] is the study of genetic material that is extracted directly from the environment. It is sometimes also referred to as environmental genomics. There are a couple of different strategies that has been employed here. In the earliest studies marker genes were targeted that are good for assessing biodiversity, especially the 16S rRNA. But as the cost for sequencing has dropped complete sequencing of entire samples has become more and more common. We knew already that we were missing quite a lot of the microbial life in nature and with this method we can identify areas that are particularly undersampled.

Global Ocean Sampling

One of the biggest metagenomic projects is the Global Ocean Sampling (GOS). In this study the marine environment was sampled at 41 sites worldwide[26]. To get an appreciation of the scale of this initiative we can only look to the fact that this dataset doubled the size of the NCBI sequence database. We have used this dataset in several of our studies, and in quite different ways as well. It is a very good resource when studying the diversity of the oceans.

1.3.2 Single-cell Genomics

Another technique, that is much more recent, is the advent of single-cell genomics. Here the genome of a single cell is amplified so that it can be sequenced. This is a great development since previously one had to grow a fairly large culture to harvest enough genetic material for sequencing. The technology that made this possible is called multiple displacement amplification[8].

The main drawback in a metagenomic study is that we can only recover one or a few genes together. Single-cell genomics is an exciting development in light of this, giving us almost entire genomes.

1.4 This thesis

During my PhD work I have focused on the evolution of the Alphaproteobacteria with specific focus on the SAR11 group. The first two papers focus on the evolution of the mitochondria. In both of these my contribution have been mainly to facilitate different bioinformatic analyzes. In Paper III and Paper V I have shown how sensitive phylogenetics can be to certain artifacts and have reinterpreted the development of specifically the SAR11 in light of this. Paper IV looks at the evolution of freshwater bacteria from the SAR11 clade and make comparisons between these and their oceanic relatives. Paper VI is a reanalysis of a previous study[38] where I believed that the result was an artifact of not taking the relationships between the taxa into account during the statistics. And finally in Paper VII we revisit the Alphaproteobacterial genome evolution with new methods where we compare structure-based orthology methods with sequence based.

2. Results

2.1 Origin of Mitochondria

In the first two papers of this thesis we have looked specifically at the evolution of the mitochondrion.

Paper I investigates the origin and evolution of the mitochondrial aminoacyl-tRNA synthetases (aaRS). The aaRS are well conserved and ubiquitous which make them good candidates for phylogenies. They are found in two copies in the nuclear genomes of eukaryotes, one that is targeted for the mitochondria and one that is targeted for the cytoplasm. In the study we used a set of 20 Alphaproteobacteria, 10 eukaryotes and another 40 representatives of other bacteria and archaea. We used the cohesion of the Alphaproteobacteria to judge whether a gene had good phylogenetic signal or not. For 12 of the 20 aaRS the Alphaproteobacteria was monophyletic, and all but one had bootstrap support above 85% for the root of the Alphaproteobacteria. In none of these 12 trees the aaRS targeted for the mitochondria clustered with the Alphaproteobacteria. We did several different tests to validate the phylogenies but could find no reason to doubt this placement. Instead we are left with the puzzle that while the respiratory chain proteins support an Alphaproteobacterial origin of the mitochondria the mitochondrial aminoacyl tRNA synthetases do not.

In paper II we tested the hypothesis that the mitochondria is closely related to the SAR11 clade. This is prompted by the suggestion that the SAR11 clusters at the base of the Rickettsiales[35] which also is the common placement for the mitochondria. This suggests that the SAR11 clade is important for our understanding of the origin of the mitochondria. With the help of the Global Oceanic Sampling (GOS)[26] we decided to investigate this further.

We used the mitochondrial genome of the protist *Reclinomonas americana* as a starting point, selecting only genes that had both the mitochondria and the Alphaproteobacteria as monophyletic clades. The 10 genes that satisfied this were then scrutinized in greater detail using bayesian methods and were also tested for sequence heterogeneity. In the end we retained 3 genes, COX1, COB and NAD7, that showed almost no AT/GC-heterogeneity and that had high support in the deeper lineages. We then used the GOS data to increase our taxonomic sampling of oceanic bacteria to see if we could shed some light on the relationship between SAR11 and mitochondria.

We could not find any support for the grouping of SAR11 and the mitochondria in any of the three proteins. But we did identify a group of sequences, distinct from the SAR11-clade, forming a group at the base of the Rickettsiales.

2.2 Alphaproteobacterial phylogeny

The first phylogenies of the SAR11 group of bacteria suggested that it is closely related to the Rickettsiales and the mitochondria[35]. This seems natural since both groups have very reduced genomes. Placing the SAR11 with the Rickettsiales would mean that we have in total two genome reductions during the course of Alphaproteobacterial evolution, the other being in the evolution of *Bartonella* and *Brucella*.

But when looking more closely at the Rickettsiales-SAR11 connection there are a number of questions that arise. The features of the genomes are wildly different for example. SAR11 have very compact genomes with few, if any, pseudogenes while the Rickettsiales have large portions of non-coding DNA[2]. They inhabit very different ecosystems, Rickettsiales being intracellular while the SAR11 can be found in oceans and lakes.

In paper III we decided to look more closely at this connection and compare the reductive processes between these two clades.

An assumption when inferring phylogenies is that the sequences have similar nucleotide or amino-acid frequencies. This is a potential problem when looking at the Alphaproteobacteria since there are big differences in nucleotide compositions between the species, ranging from 35% AT in some free-living bacteria to 70% AT in *Ca.* P. ubique (Paper III). When inferring phylogenies without taking this into account the most common result is that the SAR11, which are high in AT, group together with the Rickettsiales, which also have a high AT content [35, 32, 13].

In this paper we assembled a dataset of 71 Alphaproteobacteria. We used all-against-all blast and tribeMCL to get the orthologous proteins of the Alphaproteobacteria. Later we also extended the dataset to include more species. For our phylogeny we choose to include only proteins that were present in exactly one copy in each of the species, called pan-orthologs, since these are more likely to have been inherited vertically. This gave us a set of 58 pan-orthologous proteins for our tree. When inferring phylogenies on this dataset we got different placements for *Ca.* P. ubique dependent on whether we used RAxML or phylobayes with the CAT-model. We suspected that *Ca.* P. ubique had been artificially grouped with the Rickettsiales since they have similar AT-content.

The approach then is to try and reduce the effect of this bias. The simplest way to do this is to select genes that are less affected by the bias. We chose to use the aminoGC-measure of heterogeneity, which is the frequency of codons encoded exclusively by G or C in first two positions. The phylogenies inferred when using the proteins with the least spread in aminoGC place the *Ca.* P. ubique among the free living Alphaproteobacteria in both the maximum likelihood and in the bayesian analyzes. But when the proteins with the most spread is used, the *Ca.* P. ubique group with the Rickettsiales, even in the bayesian analysis. This is a strong indication that the previous placement

is an effect of compositional biases, but it could also be the case that it is an effect of the genes chosen. It might be that the proteins with the lowest spread had been horizontally transferred or that there is some other complication generating this difference. To address this we also developed a different method of dealing with the bias. By filtering out sites that contribute to alignment heterogeneity we were able to use all of the 58 proteins in our analysis. Also with this method we find that when using the least biased dataset the *Ca.* P. ubique is placed with the free living Alphaproteobacteria and when using the most biased dataset they instead gets placed at the root of the Rickettsiales. We also extended the dataset to include 135 Alphaproteobacteria and redid all of the above analysis with very similar results, see Figure 2D in paper III. The conclusion we draw from this is that the placement of *Ca.* P. ubique together with the Rickettsiales is an artifact of the similar mutational biases affecting the genomes in these organisms.

At the same time as we published paper III, Thrash et al. [32] published a large phylogenomic study on the placement of the SAR11. In that paper they varied the underlying dataset in several ways. They varied the set of taxa, their criteria for choosing orthologous groups and also how they treated the resulting alignments. At first sight this looks very rigorous. But we were concerned that they didn't do anything to asses the influence of compositional bias on their result. They dismiss the whole issue by pointing to their figure 7 and claiming that there is a spread of different GC composition throughout the tree. We were a bit surprised by this claim since we think that that figure shows that almost all low-GC taxa are grouped together with only two outside of the Rickettsiales clade. We were also concerned that they could only identify between 15 and 18 orthologous groups present in all Alphaproteobacteria in their most strict setting. In our analysis we had identified 58 panorthologs, though we did use fewer species, they did allow for 10% missing data in this most strict setting, so they should have recovered more.

This prompted us to reinvestigate this issue again, and also gave us a chance to also include mitochondria in the analysis. Paper V is the fruit of this initiative. We also got a chance to investigate the relative placement of the different species within the SAR11 clade.

In paper V a larger number of Alphaproteobacteria was selected from the JGI website. We acquired a set of 135 Alphaproteobacteria complemented with 8 outgroup species for rooting. We clustered the proteomes of the Alphaproteobacteria with orthoMCL. In total we found 74 proteins that occurred exactly once in all of the 135 genomes, again this in contrast to only 18 that was found in Thrash et al. [32]. The differences are probably due in part to different choice of clustering method. Where they used standard tribeMCL we had used orthoMCL. I also think that the usage of orthoMCL is the reason we found more panorhtologs than in paper III. By relaxing the criteria for orthologous groups, allowing for a few missing species and some duplications and then careful examination of each of these proteins we identified a total of

209 proteins existing in almost all Alphaproteobacteria. We then used a discordance filter[36], to remove orthologs that showed signs of horizontal gene transfer. With a final set of 150 proteins we inferred the Alphaproteobacterial species tree. Again we were able to place the SAR11 with high confidence together with other free-living Alphaproteobacteria and away from the Rickettssiales clade. Additionally the placement of HIMB59 was not inside the SAR11 and instead with the SAR116 clade.

This result also held when the mitochondria was included in the analysis. We included a set of 48 mitochondrial genomes which we clustered with orthoMCL and added nuclear encoded sequences when a protein was missing. After merger of the mitochondrial orthologs with the Alphaproteobacteria orthologs we did very careful quality checks to exclude paralogous sequences such as for example plastid proteins.

When concatenating only those proteins which gave at least 70% bootstrap support for the mitochondrial clade we got a phylogeny where the SAR11 grouped together with the free-living Alphaproteobacteria and the mitochondria with the Rickettsiales. The strict criteria we used for selecting proteins should filter out fast-evolving sequences, thus the sequences we retain should be good candidates for resolving deep phylogenetic relationships. It is important to note here that to obtain this phylogeny we didn't have to use any specific bias-filtering procedure and could even use RAxML, which was more sensitive to the biases in paper III, to infer the phylogenies.

In conclusion we find no support for grouping the SAR11 together with the mitochondria nor with the Rickesttsiales.

2.2.1 HIMB59 placement

Further arguments for placement of HIMB59 together with the other SAR11 species were put forward by Grote et al. in [17]. This paper, that mainly focus on comparative genomics of the SAR11 clade, make the following claim in support of placing HIMB59 together with SAR11 "the conservation of gene content, synteny, and the HVR2 region across these strains provides additional evidence for the shared common ancestry of HIMB59 and other SAR11 strains". We reanalyzed each of these claims in paper V. Briefly our chief method for evaluating this was to contrast comparisons between HIMB59 and SAR11 with comparisons between HIMB59 and the other Alphaproteobacteria - something that was missing in the original paper. Though they do make some comparisons, they are not specifically with other closely related species.

Gene content

Their argument is that the relative size of the core-genome in each of the species is similar to what it is for other orders.

We, instead, looked at proteome overlaps between all pairs of Alphaproteobacteria. The idea being that species that belong to the same clade should

be more similar to each other than to other species. We divided the number of shared orthologous groups with the number of orthologs in the smallest genome.

To the exclusion of HIMB59, all SAR11 species are more similar to each other then they are to any other Alphaproteobacteria. In striking contrast, HIMB59 is at best the 63:rd, out of 135, most similar species to the SAR11. Likewise when comparing HIMB59 with the other Alphaproteobacteria. The different SAR11 species ranks between 64 and 79.

We find no support for the claim that the gene content of HIMB59 should be more similar to the SAR11 than anything else.

Synteny

Here they used the ruby synteny finder[37] to calculate synteny and amino-acid identity (AAI) between the SAR11 genomes. These were then compared to other values previously published[37], arguing that the amount of synteny is similar to that when comparing synteny within other orders.

We used the same program to recalculate the synteny values. But instead of comparing to previously published figures, we also calculated synteny and AAI between every pair of Alphaproteobacteria.

The synteny we observed between the different SAR11 species, except for HIMB59, was among the highest of all the comparisons performed (see figure 4 of paper V). The comparison with HIMB59 was not so extreme. HIMB59 compared to the SAR11 species is on the higher end for gene-order conservation for HIMB59 though there are other non-SAR11 taxa that compare equally high to SAR11. Though there are species that belong to the same order that have much lower synteny values, there is also plenty of species that belong to completely different orders that have a higher gene order conservation. Casting doubts over the value of using this measure in determining whether two species belong to the same order.

HVR2

The final argument for including HIMB59 among the SAR11 is the presence of the genomic island HVR2. HVR2 is in recognized by the fact that it is flanked by 16S rRNA, $tRNA^{Ile\text{-}GAT}$, $tRNA^{Ala\text{-}TGC}$, 23S rRNA on one side and 5S rRNA on the other. Except for in HIMB59 where all of those genes can be found together in an operon in a different part of the genome. Instead they have identified a genomic island that is bounded by $tRNA^{Ser\text{-}GGA}$ and $tRNA^{Ala\text{-}GGC}$ and located at approximately the same distance from the *dnaAN* locus. They also write that the island contain the same type of genes.

We compared the sequence similarity between the different SAR11 taxa to verify that this was HVR2 in all of the species. We used tblastx here to maximize our sensitivity. In general there was some, albeit low, similarity in HVR2 between the different SAR11 species (figure 5). In HIMB59 it was almost non-existent.

Since this region is not flanked by the same genes and show almost no similarity to HVR2 in any other SAR11 our conclusion is that it is very uncertain whether this is HVR2 or not.

No evidence for HIMB59-SAR11 connection
We could not corroborate any of the claims made in favor of including HIMB59 in the proposed *Pelagibacterales* order. Neither in phylogenies nor in any of the different genomic features used as argument for its inclusion.

2.2.2 Freshwater SAR11

We have also investigated the phylogenetic affiliation of the freshwater representative of the SAR11, called LD12, in paper IV. Here we had assemblies from 10 different cells of LD12. We were interested in how these related to each other and their position relative to the other SAR11 species.

The phylogenies were done by choosing the same set of proteins that had been used in Paper III, though one of the 58 proteins were missing in all of our 10 LD12 genomes. Since we were only interested in the placement relative to the SAR11 group we only choose 4 Alphaproteobacteria as the outgroup. The obtained phylogeny showed that LD12 was monophyletic and the placement is consistent with earlier phylogenies based on the 16S gene[40]. We could identify 3 distinct subclades which was further confirmed by pairwise dS values.

2.3 Genome evolution

A further theme in this thesis is the investigation of gene gain and loss patterns in the Alphaproteobacteria.

2.3.1 Genome content

We used the distribution of orthologs to investigate the phylogenetic placement of *Ca.* P. ubique in a different way as well. We looked at clusters shared between *Ca.* P. ubique and the Rickettsiales to the exclusion of the other Alphaproteobacteria and also clusters shared exclusively between *Ca.* P. ubique and the non-Rickettsiales Alphaproteobacteria. Our reasoning here was that if *Ca.* P. ubique was more closely related to the Rickettsiales we should be able to find genes that are specific to those species, a kind of synapomorphic trait. We managed to identify 7 proteins shared between *Ca.* P. ubique and Rickettsiales to the exclusion of all other Alphaproteobacteria, though none of these were present in all, or even most, of the Rickettsiales. Furthermore their homology were quite week. The other dataset however was quite a bit

larger, we identified 372 proteins that *Ca.* P. ubique shared with the other Alphaproteobacteria. While I would not use this as proof for the placement of *Ca.* P. ubique together with the free-living Alphaproteobacteria, it is a fact that is easier to accommodate with that placement.

2.3.2 Repair and recombination

Ca. P. ubique have lost several genes involved in DNA replication and repair. Among these, 12 genes that were conserved in all other Alphaproteobacteria. To make sure that these were missing from the SAR11 clade and not only in this specific species, we checked for these in the other sequenced genomes in this clade. We managed to identify only 4 in the genome of the coastal strain HIMB114, but all were absent from both of the open ocean strains HTCC1002 and HTCC7211. Further we tried to identify *mutS* and *mutL*, both of which were missing in all of the SAR11, in the GOS metagenome. These genes were clearly under-represented when compared to genes of similar lengths that were present in the genomes, implying that these genes are really absent from this clade.

The loss of all these genes involved in recombination and repair processes could be the key to explain several of the unique features of the SAR11 genomes. Loss of repair genes is a feature associated with hypermutators[24] so this would give *Ca.* P. ubique a way to quickly adapt to different environmental conditions. Coupled with the high recombination rate SAR11 then also has a method for quickly spreading innovations through the population. The low GC-content of the SAR11 could then be an artifact of selection for adaptability instead of being something that is directly selected.

2.3.3 LD12

We compared freshwater SAR11 and the marine SAR11 in paper IV. Since each of the SAG assemblies were incomplete we decided to merge them into one pseudo-taxon. All protein clusters that were present in at least two different SAGs were included in this taxon and the tree we used for ancestral reconstruction was the same as that in paper III, although we replaced the *Ca.* P. ubique taxon with the phylogeny of the SAR11 from paper IV and we merged all the LD12 SAGs into one branch. The ancestral reconstruction showed 130 losses and 238 gains on the branch leading to LD12. Of the gains 109 could not be identified in any other Alphaproteobacteria in the study. Many of these gains could be identified to belong to a genomic island similar to HVR2 in the other SAR11. This part of the LD12 genome stands in stark contrast to the rest of the genome in that it has a very low-conservation both in terms of protein content and synteny. Furthermore we could not identify the repair

and recombination systems that have been lost in the other SAR11, so these systems were probably lost in the ancestor to the entire SAR11 clade.

Recombination rates
The oceanic SAR11-clade is exceptional in that it has very high recombination to mutation rates. It has been estimated to 63[33], which is at the upper extreme for bacteria[34]. For the LD12 SAGs we calculated this value to be 0.14, which instead is at the lower extreme. In paper III we speculated that the loss of genes involved in the recombination and repair could explain the extreme recombination rate. But this casts that hypothesis into question since also LD12 have lost these genes. It could be that the different environments, open ocean versus small lakes, is the explanation for this difference.

2.3.4 Structure-based orthology

In Paper VII we have compared sequence and structure based orthology predictions. Three different homology measures were used, the Blast based orthoMCL, HMM-based pfam and the highest level was the structure based Superfamilies. In the latter HMMs are used to assign a protein or parts of a protein to a specific Superfamily. Another difference between the three methods is that orthoMCL is an unsupervised method that determines the clusters without any supporting information while both pfam and the superfamily approach have a database of predetermined clusters. Each of these orthology predictions are expected to result in higher and higher levels of abstractions and that is indeed what we find.

We used the proteomes of a set of 71 Alphaproteobacteria. For the orthoMCL clustering we get 19157 clusters while we get 1087 superfamilies. The orthoMCL clusters also covered a larger fraction of the proteins in the dataset, 86% compared to 67% for the Superfamilies. It might be prudent to point out here that the 14% not clustered by orthoMCL is the proteins for which no ortholog could be detected.

All of these orthologous groups were mapped to a phylogeny of the Alphaproteobacteria using parsimony. We found that 739 of the 1087 identified superfamilies were present in the ancestor. These ancestral superfamilies represent the bulk (97%) of the proteins with assigned superfamilies. And then only a very small fraction of these ancestral superfamilies is responsible for these, mostly proteins involved in metabolism but also in regulation and intracellular processes. This trend that only a few of the SFs are responsible for the bulk of the proteins gets only stronger when considering the non-ancestral SFs. These correspond to one third of the number of SFs but only 3% of the proteome in each species.

The singletons that were not clustered by orthoMCL represent 14% of all proteins. Of these 3% have a superfamily assignment. We tried to deter-

mine whether these were true orfans by blasting them against the NCBI non-redundant database. For one third of the orfan proteins we could not determine any sequence or structure homolog. As the origin of these genes is highly puzzling we investigated whether any of these could be classified as Intrinsically Disordered Proteins (IDPs) by looking in the D2P2 database[23]. Almost half of our orfan proteins could be classified as IDPs. Further, the proteins classified as IDPs were shorter on average than the orfans not classified as IDPs. If these predictions are correct we have identified a way forward for determining the more exact role of a large part of the hitherto mysterious orfans.

2.4 Highly abundant bacteria

In paper VI we did a reanalysis of a study by Yooseph et al. [38]. In that study they tried to identify genes that were differentially occurring in highly versus lowly abundant bacteria. As the source dataset for this they had a set of 197 genomes from different oceanic bacteria and archaea. The abundance were estimated by mapping the reads from the GOS dataset to the genomes. They could then identify 568 protein families that were differentially expressed between the two groups. The reason we were interested in redoing this analysis was that several of the genomes in the highly abundant group were from very few species, almost half of the 34 highly abundant taxa were from *Ca.* P. ubique, *Synechococcus* sp. and *Prochlorococcus marinus*. This could mean that presence or absence in these three species would be what differentiated these groups and not something pertaining to the relative abundance of the different species.

To redo this we collapsed genomes based on the genus level and redid their statistical analysis. The cutoff between the high and low recruiting groups were decided based on how often a split at that point resulted in significantly different groups for each of the protein families in the dataset. Using the same cutoff as in the original paper we found only 2 protein families that were differentially expressed. If we instead tried to find a cutoff that generated more differentially occurring families between highly and lowly recruiting genomes, we could identify up to 50 families that differed between them. In other words, the fact that so many of the taxa in the abundant set were from so few species seems to have had a large effect on the analysis.

3. Conclusions and Future perspectives

In this thesis I have shown how difficult it can be to do phylogenetic inferences. It is very easy to forget to do quality control of datasets, especially in an environment where "big data" and "high-throughput" is fashionable. I have fallen victim to this myself during my research. When doing the analysis for what later became paper III I was very sloppy when aligning and didn't do any post processing of the alignments at all. I just took the output of the alignments and put them into the phylogenetics program. The result of this analysis was that we got *Ca.* P. ubique placed at the base of the Rickettsiales in both the maximum likelihood and the bayesian analyzes. The reason we identified that the placement of *Ca.* P. ubique could be an artifact of the AT-bias was that I decided to run the bayesian analysis for a little while longer. And lo and behold in one of the runs the placement changed. This prompted us to be more careful and especially I started to be a lot more careful when running different software. This I think, is the biggest lesson for me. In retrospect it is an obvious lesson, but apparently this was a mistake I had to make.

There is one method dealing with sequence heterogeneity that I would like someone to do with regards to the placement of SAR11. In Roure et al. [25] the authors use CAT-model of phylobayes to look at how different sites in different preselected clades get assigned different rate-matrices. By removing sites that get very different rate-matrices they were able to compensate for the heterogeneity. It would be interesting to do this here. When doing this in the Alphaproteobacteria one could compare the Rickettsiales with a some of the free living bacteria, excluding SAR11 completely from the filtering step. Preferably then one is only removing sites that are prone to be affected by heterogeneity without risking being biased towards selecting sites that are similar between SAR11 and free-living bacteria. This is a pretty involved method, but one I have great confidence in.

The exact position of the SAR11 clade is not definite in any of my analyzes. While I think I can say with high confidence that it should not be at the base of the Rickettsiales, I am unsure of the exact placement. It would make sense if it grouped together with other oceanic Alphaproteobacteria, but that's not necessarily the case. Once we get more genomes sequenced along the branch leading to the SAR11 maybe we can become more confident in where it belongs in the tree of life. The single cell-genomics approach looks very promising here.

I have used parsimony to reconstruct ancestral genomes. The main drawback of this is that parsimony does not take branch-lengths into account. Preferably one should make trees of all orthologs and then reconcile these with the

species tree. But lacking that, I think one should use likelihood or bayesian methods to estimate gains and losses on a tree. But that is also more complicated as it does not give a simple yes/no answer to the question of whether a protein is present or not in an internal node of a tree. There are a few recent developments in this area that show promise, especially the two programs Gloome[6] and Count[7] look promising. If I had another year or two I would try to use these.

4. Svensk sammanfattning

I haven lever en av världens mest framgångsrika organismer. Den är så liten att den inte går att se med blotta ögat, men det har uppsakttats att det finns totalt 10 miljarders miljarders miljarder celler från den här arten. Den här bakterien går under namnet SAR11 och har sin evolutionära hemvist hos Alphaproteobakterierna.

I alla våra celler finns det ett litet kraftverk som producerar det mesta av vår energi. Detta kraftverk förekommer hos allt flercelligt liv och i mycket av det encelliga livet. Det här kraftverket kallas för mitokondrie. Vi tror att ursprunget till mitokondrien är att en liten bakterie blivit uppslukad av en annan cell, och att den lilla bakterien genom årmiljonerna reducerats till en liten maskin. Det är inte orimligt att den lilla bakterien tillhör gruppen Alphaproteobakterier.

Alphaproteobakterier finns i de flesta miljöer. Som redan nämnts så finns de i vatten, men de finns också i jord där de bland annat förser baljväxter med kväve. Det finns också flera som kan orsaka sjukdomar hos människor, bland annat fläck tyfus. De sjukdomsframkallande bakterierna tillhör släktet Rickettsia, och det är även här som mitokondrierna brukar placeras i evolutionära anlyser.

Frågan är, har de här två något med varandra att göra. Är världens mest framgångsrika bakterie närbesläktad med kraftverken i våra celler? Flera studier har tidigare visat på det. Det är en av frågorna som behandlas i den här avhandlingen.

Jag har kunnat visa på att släktskapet mellan SAR11 och Rickettsiales uppstår på grund av att de använder nukleotiderna, A, T, C och G, i liknande proportioner i sin arvsmassa. När man tar hänsyn till detta i släktskapsanalyser hamnar SAR11 istället vid andra frilevande Alphaproteobakterier. Det finns en mängd genomdata tillgänglig som är extraherad direkt från haven. Vi har även använt denna för att ta reda på om det finns något stöd för att mitokondrien skulle vara närbesläktad med SAR11. Men vid nogrann efterforskning kan vi inte hitta någon sådan koppling.

Vi har också identifierat att en mängd gener som är viktiga för reparation av arvsmassan saknas i SAR11. Detta gör att mutationer kommer att vara vanligare än hos andra bakterier, detta skulle kunna vara förklaringen till den skeva nukleotidfördelningen.

SAR11 förekommer mest i haven, men de har även identifierats i sötvatten. Den här undergruppen går under benämningen LD12. I en av studierna i den här avhandlingen fick vi chansen att sekvensera genomen från 10 stycken olika

LD12-celler. Vi fick då den spännande möjligheten att jämföra närbesläktade bakterier som anpassat sig till olika miljöer. Den största skillnaden mellan dem är i hur stor utsträckning de utbyter genetiskt material med varandra inom grupperna. Den här typen av utbyte kallas för rekombination. De havslevande SAR11 cellerna utmärker sig för att de rekombinerar så väldigt mycket. Medan den sötvattenslevande LD12 gruppen har nästan ingen rekombination alls.

Vi har även jämfört olika metoder för att identifiera genfamiljer. Dels metoder som är baserade på sekvenslikhet och dels på metoder som är baserade på struktur. I alla genom brukar man alltid hitta ett antal gener som är olika allt anna som hitills identifierats, dessa gener brukar gå under namnet "orfans", eftersom de är föräldralösa. Vi har identifierat att minst hälften av de här generna är oordnade proteiner. Ett oordnat protein är ett protein som inte har någon struktur i sig själv utan får det det i kontakt med andra protein. I och med detta har vi identifierat en potentiell funktion för de här proteinerna.

Den största delen av arbetet i den här avhandlingen har varit att försöka hantera skevheterna i DNA för att kunna göra evolutionära analyser. Förhoppningsvis kommer det bättre modeller och program i framtiden för att hantera sådana här skevheter.

5. Acknowledgments

Först och främst vill jag tacka **Siv** för att jag fick den här chansen. Det har varit väldigt roligt och stimulerande. **Hans-Henrik** du öppnade mina ögon för vikten av databaser. **Micke**, för all kunskap om fylogenier. **Kasia**, you prepared the path for me, these last weeks, it has been invaluable. **Jonas**, hur många dragspel har du stulit egentligen? **Björn B**, du har lärt mig hur man jagar stadsduvor. **Håkan**, nu vet jag att MUDs är fitnesshöjande. **Björn N**, du har en förmåga att alltid ställa *rätt* frågor. Och så **Eva** förstås, vinner du fortfarande lika mycket? **Mats** och **Olga**, tack för att jag fick träffa er lilla **Björn** (han är inte så liten längre). **Kirsten**, you're next (I have some tips on what not to do)! **Minna**, it's so fun to follow the adventures of your little animalito. **Daniel**, I can not decide on whether you're from Spain or live in Sweden. **Mayank** may you always find good indian food. **Lionel**, din organisatoriska förmåga är svårslagen. **Robert** för alla historiska utblickar. **Ajith**, the provocateur extraordinare. **Erik** för fredagsslipsen, en fantastisk tradition. **Wei**, be careful when you delve deep into the R labyrinth, some never make it out again. **Ann-Sofie** och **Kristina**, ni förser oss med det viktigaste av allt. **Jessin**, for teaching us about the "common house lizard". Administrationen är jag djupt tacksam för, och då särskilt **Karin**, **Staffan** och **Anders**. **Jennifer** for organizing the Lennart Lottery. **Lisa** för alla kaffe och pratstunder. **Jan** för fylogenetisk nyansering. **Feifei**, god morgon! Har du fått något bollhav ännu **Rolf**? **Thijs**, for being a great scientist. **Anders**, strongt att byta doktorandprojekt. **Joran** for being so polite. And **Jimmy** and **Anja**. **Jenny**, kan jag, så kan du. **Alexander**, tack för Attwenger. Scilife-**Pall** och -**Johan**, det är bra att någon ser till att det alltid finns kaffe. And all other former patrons of molev, **Carolin**, **Magnus**, **Alistair**, **Sofi**, **Ester**, **Gustav**, **Robin**, **Otto**, **Hillevi**, **Alexandra**, **Christina**, **Ola**, **Jocke**, **Fredrik**. And also, thanks to all students I have met, you have made my work so much more fun. And to everyone else that I might have forgotten.

Emmeli, vad har inte du stått ut med de här sista veckorna, om jag ändå kunde återgälda det. Och så, till sist, mina barn, **Alva** och **Albin** utan er är jag ingen.

I love you all.

References

[1] Rudolf I Amann, Wolfgang Ludwig, and Karl-Heinz Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169, 1995.

[2] Siv GE Andersson, Alireza Zomorodipour, Jan O Andersson, Thomas Sicheritz-Pontén, U Cecilia M Alsmark, Raf M Podowski, A Kristina Näslund, Ann-Sofie Eriksson, Herbert H Winkler, and Charles G Kurland. The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature*, 396(6707):133–140, 1998.

[3] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[4] Samuel Blanquart and Nicolas Lartillot. A site-and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, 25(5):842–858, 2008.

[5] Feng Chen, Aaron J Mackey, Jeroen K Vermunt, and David S Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one*, 2(4):e383, 2007.

[6] Ofir Cohen, Haim Ashkenazy, Frida Belinky, Dorothée Huchon, and Tal Pupko. Gloome: gain loss mapping engine. *Bioinformatics*, 26(22):2914–2915, 2010.

[7] Miklós Csűös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, 2010.

[8] Frank B Dean, John R Nelson, Theresa L Giesler, and Roger S Lasken. Rapid amplification of plasmid and phage dna using phi29 dna polymerase and multiply-primed rolling circle amplification. *Genome research*, 11(6):1095–1099, 2001.

[9] Victor V Emelyanov. Mitochondrial connection to the origin of the eukaryotic cell. *European journal of biochemistry*, 270(8):1599–1618, 2003.

[10] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.

[11] Victor J Freeman. Studies on the virulence of bacteriophage-infected strains of corynebacterium diphtheriae. *Journal of bacteriology*, 61(6):675, 1951.

[12] Francis Galibert, Turlough M Finan, Sharon R Long, Alfred Pühler, Pia Abola, Frédéric Ampe, Frédérique Barloy-Hubler, Melanie J Barnett, Anke Becker, Pierre Boistard, et al. The composite genome of the legume symbiont sinorhizobium meliloti. *Science*, 293(5530):668–672, 2001.

[13] K. Georgiades, M. A. Madoui, P. Le, C. Robert, and D. Raoult. Phylogenomic analysis of Odyssella thessalonicensis fortifies the common origin of Rickettsiales, Pelagibacter ubique and Reclimonas americana mitochondrion.

PLoS ONE, 6(9):e24857, 2011. [PubMed Central:PMC3177885] [DOI:10.1371/journal.pone.0024857] [PubMed:21957463].

[14] S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245, Aug 2005. [DOI:10.1126/science.1114057] [PubMed:16109880].

[15] Stephen J Giovannoni, Theresa B Britschgi, Craig L Moyer, and Katharine G Field. Genetic diversity in sargasso sea bacterioplankton. 1990.

[16] Julian Gough, Kevin Karplus, Richard Hughey, and Cyrus Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of molecular biology*, 313(4):903–919, 2001.

[17] Jana Grote, J Cameron Thrash, Megan J Huggett, Zachary C Landry, Paul Carini, Stephen J Giovannoni, and Michael S Rappé. Streamlining and core genome conservation among highly divergent members of the sar11 clade. *MBio*, 3(5), 2012.

[18] Jo Handelsman, Michelle R Rondon, Sean F Brady, Jon Clardy, and Robert M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–R249, 1998.

[19] Nicolas Lartillot, Thomas Lepage, and Samuel Blanquart. Phylobayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288, 2009.

[20] Nicolas Lartillot and Hervé Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109, 2004.

[21] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.

[22] Robert M Morris, Michael S Rappé, Stephanie A Connon, Kevin L Vergin, William A Siebold, Craig A Carlson, and Stephen J Giovannoni. Sar11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810, 2002.

[23] Matt E Oates, Pedro Romero, Takashi Ishida, Mohamed Ghalwash, Marcin J Mizianty, Bin Xue, Zsuzsanna Dosztányi, Vladimir N Uversky, Zoran Obradovic, Lukasz Kurgan, et al. D2p2: database of disordered protein predictions. *Nucleic acids research*, 41(D1):D508–D516, 2013.

[24] Antonio Oliver, Rafael Cantón, Pilar Campo, Fernando Baquero, and Jesús Blázquez. High frequency of hypermutable pseudomonas aeruginosa in cystic fibrosis lung infection. *Science*, 288(5469):1251–1253, 2000.

[25] Béatrice Roure and Hervé Philippe. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, 11(1):17, 2011.

[26] Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, et al. The sorcerer ii global ocean sampling

expedition: northwest atlantic through eastern tropical pacific. *PLoS biology*, 5(3):e77, 2007.

[27] Lynn Sagan. On the origin of mitosing cells. *Journal of theoretical biology*, 14(3):225–IN6, 1967.

[28] Philip Siekevitz. Powerhouse of the cell. *Scientific American*, 197:131–144, 1957.

[29] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[30] Richard Stouthamer, Johannes AJ Breeuwer, and Gregory DD Hurst. Wolbachia pipientis: microbial manipulator of arthropod reproduction. *Annual Reviews in Microbiology*, 53(1):71–102, 1999.

[31] Roman L Tatusov, Eugene V Koonin, and David J Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[32] J Cameron Thrash, Alex Boyd, Megan J Huggett, Jana Grote, Paul Carini, Ryan J Yoder, Barbara Robbertse, Joseph W Spatafora, Michael S Rappé, and Stephen J Giovannoni. Phylogenomic evidence for a common ancestor of mitochondria and the sar11 clade. *Scientific reports*, 1, 2011.

[33] Kevin L Vergin, H James Tripp, Larry J Wilhelm, Dee R Denver, Michael S Rappé, and Stephen J Giovannoni. High intraspecific recombination rate in a native population of candidatus pelagibacter ubique (sar11). *Environmental microbiology*, 9(10):2430–2440, 2007.

[34] Michiel Vos and Xavier Didelot. A comparison of homologous recombination rates in bacteria and archaea. *The ISME journal*, 3(2):199–208, 2008.

[35] K. P. Williams, B. W. Sobral, and A. W. Dickerman. A robust species tree for the alphaproteobacteria. *J. Bacteriol.*, 189(13):4578–4586, Jul 2007. [PubMed Central:PMC1913456] [DOI:10.1128/JB.00269-07] [PubMed:17483224].

[36] Kelly P Williams, Joseph J Gillespie, Bruno WS Sobral, Eric K Nordberg, Eric E Snyder, Joshua M Shallom, and Allan W Dickerman. Phylogeny of gammaproteobacteria. *Journal of bacteriology*, 192(9):2305–2314, 2010.

[37] Alexis P Yelton, Brian C Thomas, Sheri L Simmons, Paul Wilmes, Adam Zemla, Michael P Thelen, Nicholas Justice, and Jillian F Banfield. A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. *PLoS computational biology*, 7(10):e1002230, 2011.

[38] Shibu Yooseph, Kenneth H Nealson, Douglas B Rusch, John P McCrow, Christopher L Dupont, Maria Kim, Justin Johnson, Robert Montgomery, Steve Ferriera, Karen Beeson, et al. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468(7320):60–66, 2010.

[39] P Zambryski, H Joos, Ch Genetello, J Leemans, M Van Montagu, and J Schell. Ti plasmid vector for the introduction of dna into plant cells without alteration of their normal regeneration capacity. *The EMBO journal*, 2(12):2143, 1983.

[40] Gabriel Zwart, Byron C Crump, Miranda P Kamst-van Agterveld, Ferry Hagen, and Suk-Kyun Han. Typical freshwater bacteria: an analysis of available 16s rrna gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology*, 28(2):141–155, 2002.

[41] Gabriel Zwart, William D Hiorns, Barbara A Methé, Miranda P van Agterveld,

Raymond Huismans, Stephen C Nold, Jonathan P Zehr, and Hendrikus J Laanbroek. Nearly identical 16s rrna sequences recovered from lakes in north america and europe indicate the existence of clades of globally distributed freshwater bacteria. *Systematic and Applied Microbiology*, 21(4):546–556, 1998.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1083

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.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-208441

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013