Evolution and diversification of secreted protein effectors in the order *Legionellales*

Tea Ammunét

Degree project in bioinformatics, 2018
Examensarbete i bioinformatik 30 hp till masterexamen, 2018
Biology Education Centre and Department of Medical Biochemistry and Microbiology, Uppsala University
Supervisor: Lionel Guy
Abstract

The evolution of a large, diverse group of intracellular bacteria was previously very difficult to study. Recent advancements in both metagenomic methods and bioinformatics has made it possible. This thesis investigates the evolution of the order Legionellales. The study concentrates on a group of proteins essential for pathogenesis and host manipulation in the order, called effector proteins. The role of effectors in host adaptation, evolutionary history and the diversification of the order were investigated using a multitude of bioinformatics methods.

First, the abundance and distribution of the known effector proteins in the order was found to cover newly discovered clades. There was a clear distinction between the proteins present in Legionellales and the outgoup, indicating the important role of the effectors in the order. Further, the effectors with known functions found in the new clades, particularly in Berkiella, revealed potential modes of host manipulation of this group.

Secondly, the evolution of the effector gene content in the order shed light on the evolution of the order, as well as on the potential evolutionary differences between Legionellaceae and Coxiellaceae. In general, most of the effectors were gained early in the last common ancestor of Legionellales and Legionellaceae, as further indication of their role in the diversification of the order. New effector genes were acquired in the Legionellaceae even up to recent speciation events, whereas Coxiellaceae have lost more protein coding genes with time. These differences may be due to horizontal gene transfer in the case of gene gains in Legionellaceae and loss of selection in the case of gene losses in Coxiellaceae.

Third, the early evolution of core gained effector proteins for the order was studied. Two of the eight investigated core effectors seem to have a connection to eukaryotes, the rest to other bacteria, indicating both inter-domain and within bacteria horizontal gene transfer. In particular, one effector protein with eukaryotic motif gained at the last common ancestor of Legionellales, was found in all the clades and is therefore an important evolutionary link that may have allowed Legionellales to utilize eukaryotic hosts.
How did bacteria in the Legionellales-group adapt to take over their hosts?

Popular science summary

Tea Ammunet

Much progress has been made in the ways we find and can study bacterial species. With the improvements in the scientific methods, much new data has been collected about new species and groups of previously known species. The knowledge about how all of these species have come about during evolution has fallen behind the huge amounts of new data. In order to know, how the many species in the group of bacteria including the pathogens causing Legionnaire’s disease and Q fever have come about, a set of important proteins was investigated in this thesis. These effector proteins aid the pathogenic bacteria to take over the host cells, and therefore enable them to act in inflicting diseases. If these proteins exist also in the newly discovered species, it may tell us, that these species can also act as pathogens.

It is further interesting to know, when in evolutionary time these proteins have come to exist. If they existed already in the early ancestors of this group of bacteria, they likely are very important in defining the whole group. In general, genes and proteins tend to be lost in time, as the bacteria evolve to use only certain host cells. The extent to which these effector proteins are maintained, gained and lost from this group may therefore tell us how the species have evolved through time.

The results of this thesis show, that the effector proteins are very widely spread within the group of bacteria Legionellales. Many of them are found even in the new species. Yet, not all of them are found outside this group, meaning that they have a certain function for this group of species in particular. Moreover, many of these proteins were taken up or evolved very early during evolution, when this group became distinctively different from other bacteria. This indicates that the effectors are, indeed, very important in defining the group Legionellales.
1 Introduction

Host-adapted, intracellular bacteria, such as the species in the order Legionellales, are difficult to cultivate in a laboratory setting due to their complex metabolic requirements. However, recent advances in cultivation-independent genomics (e.g. metagenomics) have made it possible to sequence environmental samples and reconstruct (almost) complete genomes from novel clades. These methods have improved our current understanding of the phylogeny of Legionellales and also revealed many previously unknown species.

The discovered diversity raised many questions both about the biology of the new species and about how the diversity has evolved. So far, representatives from six genera in the order Legionellales have been sequenced, namely Legionella, Rickettsiella, Diplorickettsia, Coxiella, Berkiella and the relatively newly discovered Aquicella. Their lifestyles vary from facultative intracellular to obligate, mutualistic insect endosymbionts (see e.g. Mittl and Schneider-Brachert 2007, Qiu and Luo 2017, Santos et al. 2003). Although many species are amoebal pathogens, and some are accidental human pathogens (Legionella pneumophila, L. longbeachae, the agents of Legionnaires’ disease, and Coxiella burnetii, the agent of Q-fever), the ecology and functions of the other species are largely unknown.

The mode of evolution in bacteria goes in general from a period of innovation and acquirement of genes to loss of non-essential genes (Wolf and Koonin 2013). For Legionellales it has been proposed, that most of the essential structural genes defining the order have been gained once, at the last common ancestor (Hugoson 2017). If so, these genes would thus define the order that branches off from other gammaproteobacteria quite early on. From there, diversification into the different clades would have taken place, possibly including a reduction in genome size.

One of the drivers of diversification is host adaptation. With the multitude of hosts and lifestyles present in the order, this kind of adaptation may well be behind the observed diversity. Evidence of host adaptation include loss of housekeeping genes and a general reduction of genome size, when bacteria evolve from a free-living extracellular lifestyle to an obligate intracellular lifestyle.

Effector proteins, used to take over the host cell functions, provide good example of a common group of essential proteins, that nevertheless show a role in host adaptation in Legionellales. In total close to 6000 effector proteins have been predicted by machine learning methods (Burstein et al. 2016). Around 330 effectors have been experimentally verified (Qiu and Luo 2017), but in total 9300 effector proteins have been suggested to exist (Burstein et al. 2016). However, only a few of the effector proteins seem to be shared even among the family Legionella, meaning that many of them may have been lost or acquired specifically for adapting to a particular host.

In general, comparing the genetic and genomic organisation of any known organism with newly discovered organisms, may give indications of both the evolution and the
ecology of the species. The importance and abundance of effector proteins makes them a good candidate for use in evolutionary analysis of an order. Therefore, in order to shed light on the evolution of the order *Legionellales*, the occurrence of the effector proteins among the species was investigated. Further, the occurrence of these proteins on an evolutionary time scale was studied. Similarity or homology to known effector proteins in the newly discovered clades may also shed light to their biology, and was thus additionally studied in this thesis.

2 Background

The infamously well known species of the order *Legionellales* are *Legionella pneumophila* and *Coxiella burnetii*. Primarily infecting amoeba, *L. pneumophila* gains virulence after amoebal infection (Swanson and Hammer 2000), and may infect human and other mammalian alveolar macrophages when inhaled. The infection by *L. pneumophila* and *L. longbeachae* may cause a fatal form of pneumonia (Legionnaire’s disease) or a milder Pontiac fever in humans (see e.g. Carratalà and Garcia-Vidal 2010, Khodr et al. 2016, Qiu and Luo 2017 for review). Similarly, *C. burnetii* can cause Q fever, a zoonosis that spreads from livestock to humans (see Ghigo et al. 2009, Maurin and Raoult 1999 for review). Q fever is regularly asymptomatic, but may develop fatal in patients with cardiac diseases, lessened immunoresponse and in pregnant women.

The pathogenicity of many bacterial species is linked to specific genes or genomic regions. In the order *Legionellales*, a specific Dot/Icm Type IV Secretion System (T4SS) and its secreted effectors are keys to the endosymbiotic and pathogenic lifestyle (e.g. Segal et al. 2005). Many *Legionellales* species manipulate their host cell functioning and behavior by injecting effector proteins in to their host cytoplasm via the Dot/Icm secretion system. In the case of *L. pneumophila*, these effector proteins make up to 10% of the genome, whereas approximately 6% of the *C. burnetii* genome consists of genes encoding effector proteins (Qiu and Luo 2017).

Recently, nearly 6000 effector proteins were predicted from 38 *Legionella* species by machine learning methods (Burstein et al. 2016). These proteins formed 608 orthologous groups, of which only seven were shared among the 38 *Legionella*. The large number and diversity of effector proteins show a great deal of redundancy. The maintained high numbers can be due to several proteins either affecting the same pathway, having been recently acquired/duplicated or playing a role in host specific and environment specific adaptation (Burstein et al. 2016).

*Coxiella burnetii* also secrete effectors into their host using a similar Dot/Icm Type IV secretion system, although its overall lifestyle differs slightly from that of *L. pneumophila*. Approximately 133 *C. burnetii* effectors have been found so far. Some of the *C. burnetii* effector proteins are similar to, but most of the effector proteins can be distinguished from those of *Legionella pneumophila* (Carey et al. 2011, Chen et al. 2010). The divergence
between the effectors of these two species, is most likely due to different developmental histories: *Legionella* have evolved to infect a variety of protozoan hosts, where as the primary hosts of *C. burnetii* are mammals (Carey *et al.* 2011). Similarly to *L. pneumophila*, redundancy of the effector proteins has been observed in *C. burnetii*. Only 16 non-plasmid encoded proteins are conserved between different pathotypes of *C. burnetii* (Van Schaik *et al.* 2013). As with *Legionella*, the small proportion of concerved genes indicates that the different proteins may have, for example, emerged due to host adaptation.

Inside the host cell, with the help of the effectors, *Legionella pneumophila* first builds a *Legionella*-containing vacuole (LCV), recruits vesicles from the endoplasmic reticulum (ER) and prevents the fusion of the symbiont-containing vacuoles with lysosomes (Qiu and Luo 2017, Rolando and Buchrieser 2014, Swanson and Hammer 2000), to name a few. The effectors in *C. burnetii* help maintain an acidic environment of the formed *Coxiella* containing vacuole (CCV) (Carey *et al.* 2011). In comparison to LCV, the CCV develops from a phagosome to a phago-lysosome, that eventually fills almost all of the host cytoplasm (Van Schaik *et al.* 2013). In contrast to *Legionella*-effectors, it seems that *Coxiella*-effectors are not directing or aiding in the formation or maturation of the CCV. However, in the process, approximately 8 hours after infection, the effectors in *C. burnetii* are secreted (Van Schaik *et al.* 2013). Both *Legionella pneumophila* and *Coxiella burnetii* effectors are important in redirecting vesicle trafficking, slowing down or preventing apoptosis (Latomanski *et al.* 2016, Qiu and Luo 2017), and some have been shown to have a connection with the virulence of the species (Shames *et al.* 2017).

Many of the effectors found in *Legionella* species and in *C. burnetii*, are similar to proteins found in eukaryotes (Chen *et al.* 2010, Gomez-Valero *et al.* 2011a, 2014, Lifshitz *et al.* 2013). Usually these effector proteins contain domains, such as ankyrin repeats, coiled coils and U-boxes, that are widely spread among eukaryotes (de Felipe *et al.* 2005, Gomez-Valero *et al.* 2014, Lifshitz *et al.* 2013). Further, the genes encoding these effector proteins may have a diverging G-C content compared to other genes in the genome (de Felipe *et al.* 2005, Van Schaik *et al.* 2013). Because of their similarity to eukaryotes, it has been suggested that they have evolved closely with the host adaptation process (Gomez-Valero *et al.*, 2011a). The hypothesis about how these proteins have been acquired, include convergent evolution from ancestrally inherited genes and inter-domain horizontal gene transfer (HGT) from eukaryotes (de Felipe *et al.* 2005, Gomez-Valero *et al.* 2014, Van Schaik *et al.* 2013). The similarity of the effector proteins to eukaryotic proteins may be essential for overtaking host cell functions using molecular mimicry (Gomez-Valero *et al.* 2014).

Simultaneously with the increasing knowledge about *L. pneumophila* and *C. burnetii* effector proteins and their functions, many novel clades within the order *Legionellales* have been discovered (Figure 1). In addition to the *Aquicella*-clade, that was described from water samples in 2003 (Santos *et al.* 2003), other unidentified sequences have been found from, for example, the TARA North Pacific Ocean samples (closely related to *Coxiella*)
and from a malaria mosquito *Anopheles gambiae* (Lionel Guy, personal communication 2017). The ecology and host range of these species is, however, still largely unknown.

It is currently hypothesized that there may have been only a single evolutionary event, where most of the *Legionellales* genes were gained; many genes, including at least two of the seven shared effector proteins were gained at this time point as well as the proteins forming the Dot/Icm secretion system (Hugoson 2017). The divergence between the effectors in *Legionella* and *Coxiella* indicate, that many, if not most of the proteins gained in early evolutionary phases, were lost or radically changed. However, if all the effector proteins were also gained in one time point, and to which extent the effectors found in *L. pneumophila* and *C. burnetii* are present in the other species and clades of the order, is not known. Furthermore, to our knowledge, the presence of the effector proteins in the novel clades has not yet been thoroughly investigated.

The presence of the effector proteins in the novel clades may reveal information about the ecology and biology of these species. In particular, the presence of the eukaryotic-like proteins may reveal potential host organisms for the recently discovered clades, such as *Aquicella*. Furthermore, similarity to proteins found in eukaryotes of these eukaryotic-like effector proteins could shed light on the evolutionary context of these particular proteins.

### 2.1 Aims

The general aim of this master thesis project, was to explore the effector proteins in the newly discovered and more well known genomes of the order *Legionellales*. In order to reveal the biology and evolution of the effector proteins in this group, the previously discovered knowledge on the novel clades and sequences in the order *Legionellales* was utilized and combined with comparative genetics methods.

In more detail, the project investigated the presence and absence of known *Legionella pneumophila*, *Legionella longbeachae* and *Coxiella burnetii* effectors within the order. Furthermore, the evolution of the gene content of the order, particularly the gains and losses of effector proteins, was investigated. In addition, the early evolution of the gained core proteins in the order *Legionellales* was examined, particularly in relation to their evolutionary history with eukaryotes.

### 3 Materials and methods

The aims and the questions of the project were answered by following the general work flow presented in Figure A.1. Most of the work was carried out with *bash* scripts and coding in *Python* and *R*. All *bash* scripts may be found in the public *Bitbucket* repository (https://bitbucket.org/evolegiolab/legionellaleseffectors). Scripts are referred to by their names in the descriptions below, and a short description of them is included in Table A.1 in the appendix. In addition, manual work was done collecting the published data and
protein blast was run online (https://blast.ncbi.nlm.nih.gov/Blast.cgi) in order to reveal the early evolutionary dynamics of the proteins.

3.1 Published material

The collection of the list of experimentally verified *Legionella pneumophila*, *Coxiella burnetii* and *Legionella longbeachae* effectors was gathered from recent research and review articles. In total 13 published articles provided information and/or an experimentally tested confirmation of 160 *Coxiella burnetii* effector proteins (Chen et al. 2010, Cunha et al. 2015, Fielden et al. 2017, Graham et al. 2015, Lifshitz et al. 2013, 2014, Weber et al. 2013), 337 *Legionella pneumophila* effectors (de Felipe et al. 2005, Gomez-Valero et al. 2011a,b, Huang et al. 2011, Lifshitz et al. 2013, Qiu and Luo 2017, Zhu et al. 2011) and 129 *Legionella longbeachae* effectors (Gomez-Valero et al. 2011a,b, Lifshitz et al. 2013). In addition, 42 potential but experimentally unverified *Legionella (Fluoribacter) dumoffii* (Lifshitz et al. 2013), 41 *Legionella drancourtii* (Lifshitz et al. 2013) and 18 *Rickettsiella grylli* (Lifshitz et al. 2013) effector proteins were found. Because the effectors of the latter three species were not experimentally verified, they were excluded from further analysis.

All of the 129 *L. longbeachae* effector proteins, and 25 of the 160 *C. burnetii* effector proteins were claimed to be homologs of *L. pneumophila* effectors. Homology was determined by the inclusion of “effector domains”, such as the ankyrin domain and the Ser/Thr kinase domain (Lifshitz et al. 2013), and/or a local alignment e-value (Chen et al. 2010, Lifshitz et al. 2013, Weber et al. 2013).

Effector protein sequences were fetched from the National Center for Biotechnology Information (NCBI) using *Entrez Direct* NCBI access provider (Kans 2013) through *bash* scripting (edirect_test.sh). The collected effector protein locus tags were listed with notes on the species and the references. Each locus tag was then searched from the NCBI protein database using functions *esearch* and *efetch*. The protein accession number was extracted from the results and used to get the protein sequence. Protein accession number and protein annotation were then extracted from the sequence result, and adjoined to the list of locus tags (combine_info.sh).

3.1.1 OrthoMCL protein groups

A recent collection of *Legionellales* genomes was gathered and assembled from metagenomics data in previous work (Hugoson 2017). This collection of sequences was annotated with *prodigal* (Hyatt et al. 2010). Homologous genes were then grouped into protein profiles with *OrthoMCL* (Li et al. 2003). *OrthoMCL* is based on all-against-all blast, after which the results are assigned to a graph with sequences as nodes and similarities as edge weights. A Markov Cluster algorithm is then applied on the graph, resulting in clusters of orthologous proteins, here onwards called protein clusters. These previous results of protein clusters were used as the basis when searching effector proteins from among
3.2 Searching effectors from predicted protein clusters

The presence of the listed effector proteins in the order *Legionellales* was explored by searching through the existing orthologous protein clusters (protacc_to_prodigal.sh). In more detail, protein accession numbers from the effector list were first used to couple the effectors to the unique sequence identifiers used in OrthoMCL. This identifier was then searched for among the protein clusters and all the sequences each cluster contains. The locus tags, species, protein accession, unique identifier and protein cluster for each locus tag were then listed in a table.

3.2.1 Homologous protein clusters

Some of the effector proteins were found in different protein clusters. Particularly finding effector proteins in different species from the same protein clusters prompted the question, whether these effectors are homologous. In order to investigate this, we studied the phylogenetic trees of these proteins. The non-unique clusters (including the potential homologs), with all the sequences included in them, were combined, aligned with MAFFT (Katoh et al. 2002, 2005, align_homologs.sh) and trimmed with Trimal (Capella-Gutiérrez et al. 2009). A maximum of 30% of gaps was allowed when trimming the aligned sequences. A phylogenetic maximum likelihood tree was constructed with IQ-Tree (Hoang et al. 2018), using the WAG (Whelan and Goldman 2001) amino-acid substitution matrix with empirical codon frequencies and gamma rate heterogeneity (trimm_n_tree.sh). The trimmed alignments and the phylogenetic trees were visually inspected in AliView (Larsson 2014) or FigTree (Rambaut 2014) in order to infer homology.

3.3 Presence of effector proteins in the order

Listed unique effector protein clusters were used to count the number of copies for each effector protein in each species in the order *Legionellales* (effector_occurrence_forAlleffs.sh). Python code with Pandas and Seaborn packages were utilized to calculate further aspects and to visualize the data (effector.table.ipynb).

In order to investigate the differences between families and smaller species groups, key numbers, such as averages and proportions, were calculated dividing the species into these groups. For the bigger species groups, the species were assigned to *Legionellaceae*, *Coxiellaceae* and the outgroup (Figure 1). The outgroup species were selected so, that it would include representative big genomes from other gammaproteobacterial families, as well as orders from betaproteobacteria. For smaller species groups the species in the *Coxiellaceae*-group were further assigned to *Aquicella*, *Berkiella*, *Coxiella*, *Rickettsiella* and the general Gammaproteobacteria bacterium-groups (Figure 1). The group *Aquicella*
was assigned very conservatively, including only species named *Aquicella* and two species further in the clade. Thus this group, as it is named here, is polyphyletic.

### 3.4 Inferring the evolution of gene content

Gene content evolution was inferred by using the program **Count** (Csurós 2010). In short, the program uses a phylogenetic tree and the occurrences of homologous proteins/genes to estimate the likelihoods of phylogenetic birth-and-death-rates. A bayesian tree constructed from 109 highly conserved single-copy genes, of the order *Legionellales* and an outgroup, was previously constructed with Monte Carlo Markov Chain sampler in **phyloBayes** (Lartillot and Philippe 2004, 2006, Lartillot *et al.* 2007) under a site specific CAT-GTR model (Figure 1, Lartillot *et al.* 2007). Also a previously constructed family size table of all protein cluster occurrences in the order was used as an input to **Count**. Optimized birth-and-death rates for each branch were calculated from the family size table in **Count**. Root family size distribution was assumed to follow a Poisson-distribution. The output gives the rates for gains, losses and duplications for each branch in the input tree. Posterior probabilities for family sizes in inner nodes was then computed based on the modeled gain- and loss-rates.

**Count** results on the number of gained and lost protein clusters was parsed with an existing **Python** code, linking the gains and losses to tree node numbers. All gained and lost clusters with a probability higher than 50%, were then listed and counted (allModels_GainLoss.sh). Further, the list of effector protein clusters was used to mark, when in the tree (node) each effector protein cluster was gained and lost, as well as how many gains and losses were observed per node (effectorSearch_fromGainLoss.sh). A combination of **R** and **Python** code was then used to a) transfer node annotations from **Count** to the phylogenetic tree in the program (tree_mods.R) and b) to visualize the number of gained and lost effector protein clusters in the *Legionellales* tree (tree_visualisation_GL.ipynb).

### 3.5 Early evolution of core gained effectors

Gained effector proteins in particular in the last common ancestor (LCA) of *Legionellales*, as well as those gained in the LCA of *Legionellaceae* and *Coxiellaceae* will tell us about the early evolution of the order. The gained effector protein clusters in these nodes were thus investigated further in relation to other organisms.

#### 3.5.1 Effector protein clusters

Sequences from the gained effector protein clusters were combined, aligned with **MAFFT --liinsi** (mergeEffectorModels.sh) and trimmed with **Trimal** (gains_trimm_n_tree.sh). A maximum of 20% gaps was allowed, when trimming the aligned sequences. This amount of gap allowance was chosen based on visual inspection. In order to compare the gained
Figure 1: A phylogenetic tree of the order *Legionellales* (coloured) and the outgroup (black). The tree was constructed with the Markov Chain Monte Carlo algorithm in phyloBayes using 109 highly conserved single copy genes under CAT-GTR model. Substitutions per time unit are presented by the scale. The family *Legionellaceae* is marked with red and the family members of *Coxiellaceae* in other colours. Futher grouping into *Coxiella* (dark blue), *Rickettsiella* (light blue), *Aquicella* (green) *Berkiella* (purple) and other *Gammaproteobacteria bacterium* (orange) are marked on the tree. The branch support values were all found to be 1.
protein clusters against other organisms, the proteins of a cluster were first combined as a protein profile (pssm-matrix). This was done by first constructing a blast database from each protein cluster. Then, **psiblast** (Bhagwat and Aravind, 2007) was run comparing each protein cluster against a database of itself (eukaryote_blast.sh). The resulting position-specific scoring matrix (pssm) was then used in a further online **blastp** run against everything else except **Legionellales**. The upper limit for accepted e-value was set to $10^{-4}$. From the blast alignments, we could see that, in some cases, only part of the protein was aligned with multiple organisms. Thus, in order not to take species specific parts of the proteins into account, only the aligned parts of the resulting blast hits were exported as fasta files for further studies.

Blast result identification lines were then modified (seq_id_fix.sh) to be compatible with previous notation and one representative sequence per species was kept. The sequences were then combined with those from the respective effector protein clusters. The combined sequences were realigned with **MAFFT --add** (Katoh and Standley 2013) and trimmed with **Trimal** (Capella-Gutiérrez et al. 2009, blast_hit_trees.sh). When trimming, 20% of gaps was allowed. Maximum likelihood trees were built using LG (Le and Gascuel 2008) general amino-acid matrix with empirical codon frequencies and gamma rate heterogeneity in **IQ-Tree** (Hoang et al. 2018, blast_hit_trees.sh). Phylogenetic trees were then visualized in **FigTree** (Rambaut, 2014).

4 Results

4.1 Effectors in predicted protein clusters

Out of the 626 locus tags found in **C. burnetii**, **L. pneumophila** and **L. longbeachae**, 497 were connected to a protein cluster. Some of the locus tags were not found from the NCBI protein database reducing the number of protein accession numbers first by 22. Further, some of the effector protein accession numbers were not found from the re-annotated genomes, resulting in 572 effector proteins for the three species. Out of the total 572 **C. burnetii**, **L. pneumophila** and **L. longbeachae** effectors, 497 were found from the orthoMCL protein clusters/profiles. Some of the effectors were found from the same protein cluster, giving us 375 unique effector protein clusters.

4.1.1 Homologous protein clusters

There were in total 36 cases, where several clusters contained locus tags, that were classified as homologous in the published papers. In 26 out of 36 cases, the phylogenetic tree showed a clear division of the clusters. (Figure 2a).

In ten cases, the sequences from the annotated homologous clusters, were more or less intertwined with each other (Figure 2b). In three of the ten cases, species that were present in the intertwined cluster were also present elsewhere in the tree. This suggests
Figure 2: Representative examples of phylogenetic trees for annotated homologous proteins. The maximum likelihood method in IQtree was used to generate the trees under WAG-substitution matrix and gamma rate heterogeneity. Maximum likelihoods with bootstrap values are presented above branches. Substitutions per time unit are given by the scale.

(a) An example of a division between annotated homologous effector protein clusters. The protein clusters refer to locus tags lpg0021 and LLO_0047 (cluster0111127, blue), CBU_0235 (cluster0110446, yellow) and CBU_0682 (cluster0112917, red).

(b) An example of potentially homologous protein clusters. The protein clusters refer to locus tags lpg2271 and LLO_2530 for cluster0112816 (blue) and LLO_1728 for cluster0119809 (red).

4.2 Presence of effector proteins in the order

The number of gene copies in each of the 359 effector protein clusters are visualized as a heatmap in Figure 3. Most of the proteins are present in one copy, but up to 15 gene copies of one protein cluster were found in one species. From Figure 3, we can see clear evidence of a possible duplication event and further development of the proteins as paralogs. In two cases, the species in the intertwined cluster were not present at all elsewhere in the tree. In these cases, it seems plausible, that a horizontal gene transfer may have taken place between species in the family *Legionella* and the family *Rickettsiella*. In five cases, the branches had very low support values throughout the tree, making the interpretation of the trees uncertain. The general trend in these trees was that one or two sequences from one cluster were intertwined with the branches of the other cluster/clusters, and exhibited bootstrap maximum likelihood values below 70 or even below 30 on these branches.

Due to the low number of potential orthologous effector protein clusters, and the uncertainty of many of the trees, the protein clusters were treated individually in the further analysis.
Figure 3: The number of effector protein gene copies present for each species in the Legionellales order, and the outgroup. Gene copy counts are marked with colour, from zero (white) to 8 or over (purple). The groups Legionellacea, Coxiellacea and the outgroup are marked with red, blue and black rectangles, respectively.

The lack of many effector protein clusters is evident in the species at the bottom of the list, forming the outgroup.

Further analysis of the average gene copy numbers per species in the bigger species groups per cluster reveal, that indeed, when some gene copies are still present in both Coxiellaceae and Legionellaceae, the average gene copy numbers for the species in the outgroup is zero (Figure 4). In total, 339 effector protein clusters are present in the Legionellaceae, that forms the basis of our set of effector proteins, making 20 effector protein clusters unique for Coxiellaceae. Within the Legionellaceae group, six effector protein clusters were shared between all the species in the group. No effector protein cluster was present in all of the species within Coxiellaceae, in contrast to the outgroup,
within which 11 effector protein clusters were present in all the species.

Figure 4: The average number of gene copies per species per cluster for families Coxiellaceae (blue), Legionellaceae (orange) and outgroup (green).

A similar trend, albeit with more variation, can be seen for the smaller species groups presented in Figure 5: in all the other groups the average gene copy number per species for some of the effector protein clusters peaks above zero for about half of the protein clusters, except for the outgroup.

The total number of effector protein clusters present per group are 84, 47, 60 and 98 for Aquicella, Rickettsiella, Berkiella and Coxiella, respectively. The Berkiella-group species shared the most effector protein clusters with 33 out of 60 clusters present in all three species. More similarly to Legionellaceae and Coxiella, the species in the Rickettsiella group and in the Aquicella group shared 19 out of 84 and 16 out of 47 effector protein clusters, respectively. The numbers of shared protein clusters within the groups for Legionellaceae, Coxiella and the outgroup are the same as above for the bigger groups.

In total, 290 effector protein clusters were not present at all in the outgroup consisting of 18 species, of these 147 are visibly lacking from the outgroup as a white area in Figures 3 to 6. The 290 clusters not present in the outgroup include approximately 65% hypothetical
Figure 5: The average gene copy number per effector protein cluster for smaller species groups: Aquicella (blue), Berkiella (orange), Coxiella (green), Gammaproteobacteria bacterium (red), Legionellaceae (purple), outgroup (brown) and Rick-ettyella (pink).

proteins, but also known Dot/Icm-secreted effectors, such as SidC, SidE, SidD, SidF, SidH, VipA, VipD, RalF and SdbC. Moreover, these missing effectors include proteins
with common eukaryotic motifs (ankyrin repeats, coiled coils).

Finally, the proportion of species per each (small) species group, where the cluster was present, was calculated. In Figure 6, the proportion is depicted as a heatmap. Again, it is evident that a group of effector proteins is fairly common among all the groups (dark blue), but almost a half of the protein families are missing from the outgroup (white).

![Figure 6: The proportion of species per group where the cluster is present. The proportions are marked by color from 0 to 5% (white) and from 5-20% (light blue) to 80-100% (darkest blue). The proportions were counted for each smaller species group.](image)

### 4.3 Evolution of gene content

Altogether 438 effectors were gained and 420 effectors were lost throughout the evolution of the order *Legionellales* at different stages and in different branches (see Figures 7, 8 and A.2). The highest number of gained effector genes (locus tags) at a single point in time was estimated to be 56 (Figure 7), and the highest number of lost effector genes (locus tags) at a single time point was estimated to 26 (Figure 8). The highest numbers of gained and lost effectors originated from 54 and 25 effector protein clusters, respectively.

Figure 7 shows the gains and losses for the *Legionellaceae* family. The red, gained, effectors are present almost throughout the tree, in both inner and terminal nodes. Most of the gained effectors, however, appear early in the tree. With time, and likely adaptation, the effectors have been gradually also lost in the branches, showed by the blue color.

The evolution regarding gained and lost effectors in the *Coxiellaceae* family is shown in Figure 8. Effectors were predicted to have been gained earlier in the evolution also for this family. Some gains appear also in recent speciation events, such as for the recently discovered *Aquicella*. However, the deeper branches and terminal nodes are, in general, dominated by gene loss due to potential loss of selection. This is visible particularly for the endosymbiotic *Coxiellaceae*.

The last common ancestors (LCA) to all *Legionellales* can be seen in Figure 8. In both, LCA, sensu lato including *Berkiella* (first node from the left) and sensu stricto (second node from the left), we can observe a few gained effectors. These core gained
Figure 7: Modified phylogenetic tree from Figure 1 of Legionellaceae showing the gained (red) and lost (blue) effector proteins for each node. If no effector was either gained or lost, the node does not have a graph beside it.

Effectors, in addition to the ones gained for the common ancestor of all Legionellaceae and Coxiellaceae and their early evolution were investigated in further detail.
4.4 Early evolution of core gained effectors

All of the core effectors and information on them are listed in table 1. Two of the eight gained core effector protein clusters at the LCA of *Legionellales* showed similarity to other orders than bacteria. These were cluster0110918, corresponding to locus tag lpg2300, and cluster0111073, corresponding to locus tag lpg0896. Lpg2300 codes for an ankyrin repeat, whereas lpg0896 codes for a Sel1 protein family. These effectors were gained in the LCA of *Legionellales* sensu lato and sensu stricto, respectively.

The concatenated effector protein cluster and blast hit results tree for cluster0110918 can be seen in Figure 9. In the mid-point rooted tree, all *Legionellales* species (dark green) aggregate together on one main branch, with some other bacteria (black), a couple of fungi (yellow) and *Trichomonas vaginalis* as the only other eukaryote (blue). In the sister clade we can see more fungi, sporadic other eukaryotes, *Nicotina*-plants (pink) and the main cluster of eukaryotes consisting of mammals, fish and a snake. The other
Table 1: Gained core effectors, their nodes of gain and protein families.

<table>
<thead>
<tr>
<th>Node for gain</th>
<th>Cluster</th>
<th>Locus tags</th>
<th>Protein family</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Legionellales</em> sensul lato</td>
<td>cluster0110918</td>
<td>lpg2300</td>
<td>Ankyrin repeat</td>
</tr>
<tr>
<td><em>Legionellales</em> sensu lato</td>
<td>cluster0111236</td>
<td>lpg1565</td>
<td>NMT1 superfamily</td>
</tr>
<tr>
<td><em>Legionellales</em> sensu stricto</td>
<td>cluster0111073</td>
<td>lpg0896</td>
<td>Sel1 superfamily</td>
</tr>
<tr>
<td><em>Legionellales</em> sensu stricto</td>
<td>cluster011084</td>
<td>CBU_2076</td>
<td>putative conserved family Yqf0</td>
</tr>
<tr>
<td><em>Legionellales</em> sensu stricto</td>
<td>cluster0111097</td>
<td>CBU_0560</td>
<td>TraI_2 superfamily</td>
</tr>
<tr>
<td><em>Legionellales</em> sensu stricto</td>
<td>cluster0112144</td>
<td>CBU_0676</td>
<td>SDR superfamily</td>
</tr>
<tr>
<td><em>Coxiellaceae</em></td>
<td>cluster0112917</td>
<td>CBU_0682</td>
<td>methyltransferase_11</td>
</tr>
<tr>
<td><em>Coxiellaceae</em></td>
<td>cluster0113029</td>
<td>CBU_1334</td>
<td>ALMT superfamily</td>
</tr>
</tbody>
</table>

bacterial clades consist of, for example, *Clamydiales*-species, *Brachyspira alvinipulli* and other endobionts. The branch support values vary from 100 to 20, with main branch division likelihood estimated at 88. Thus, it is plausible, that this protein descends from a common ancestor with the eukaryotes.

Sel1 family protein tree based on effector protein cluster0111073 can be seen in Figure 10. The majority of the blast hits outside *Legionellales* were other bacteria. The bacterial hits came mostly from alpha-, beta- and other gammaproteobacteria, but also from enterobacteria, other pathogenic bacteria, such as *Massilia* and *Vibrio* and some endosymbionts. In addition, some eukaryotic species, including five mammalian species, gave positive hits with the scoring matrix of this cluster.

The *Legionellales*-species are fairly well clustered together, as expected, in one of the two main branches. One plant species and ten eukaryotic species, are all located in the second main branch.

The support values for the maximum likelihood tree varied from 6 to 100. The main branch division in this midpoint rooted tree got a low support of 28. Overall, the majority of the branch support values were low, and thus not very reliable.

The six other core gained effector protein clusters did not get significant hits from outside bacteria. After midpoint rooting all the trees, *Legionellales* species grouped well together in three of the phylogenetic trees, whereas in the other three, the species were dispersed among other bacterial species in the tree. The trees where *Legionellales* were grouped together were for cluster0111097, corresponding to locus tag CBU_0560, cluster0111236, corresponding to locus tag lpg1565 and cluster0113029, corresponding to locus tag CBU_1334. These effectors were gained in the LCA of *Legionellales* sensu stricto, *Legionellales* sensu lato and *Coxiellaceae*, respectively. cluster0111097 effector is most similar to the TraI-2 superfamily, cluster0111236 is alike the NMT1 superfamily for thiamine synthesis, and cluster0113029 is similar to the ALMT superfamily for aluminium activate malate transporter (Table 1).

The other trio of core gained effector protein clusters were cluster0111084, corre-
Figure 9: Maximum likelihood tree of combined sequences for cluster0110918 and corresponding blast hits with the e-value threshold of $10^{-4}$. The tree was built with IQtree, under LG substitution model with gamma distribution for rate variation. Bootstrap-values are shown above branches, and the scale marks the number of substitutions per time unit. *Legionellales* species are marked with dark green, other bacteria with black, fungi with yellow, plants with pink and other eukaryotes with blue color.

Corresponding to locus tag CBU_2076, cluster0112144 corresponding to CBU_0676 and cluster0112917 corresponding to CBU_0682. These clusters are similar to a putative conserved family Yqf0, the SDR superfamily for dehydratase and to SmtA methyltransferase, respectively. The clusters were gained in the LCA of *Legionellales* sensu stricto, for clusters 0111084 and 0112144, and in the LCA of *Coxiellaceae* for cluster0112917 (Table 1).
5 Discussion

In this master thesis, the evolution of the effector proteins, found in the order Legionellales, was explored by first investigating the presence of Legionella pneumophila, Legionella
longbeachae and Coxiella burnetii effectors. Secondly, the evolution of the gene content was inferred from a phylogenetic birth-and-death-rates model. Last, the early evolution of the order was studied by comparing effector proteins gained in the LCA of Legionellales to known proteins in other orders.

5.1 Effectors in the order

Overall, many of the 359 investigated effector protein clusters appear in several species in the Legionellales order (Figure 3). However, 290 protein clusters were missing from the outgroup, making a clear difference between the order and the outgroup (Figures 4 and 5). In addition, differences between groups in the order could be seen, some of which are likely due to the bias in the published effector proteins towards Legionellaceae.

Multiple copies, up to 15, of some of the effector protein clusters were found in the order Legionellales. Although redundancy in effector proteins is known, due to them functioning in the same pathway, duplication events leading to paralogs are also a potential source for divergence and new functions. There is already some indication, that paralogs of certain effector proteins could have host-specific functions (Cazalet et al. 2004, Gomez-Valero et al. 2011a).

Although none of the effector protein clusters were found in all the species in Legionellales and the outgroup, there is a small proportion of effector proteins that seem to be common in both the outgroup and Legionellales. These include both proteins linked to common functions, such as transport and metabolism, as well as effector proteins with potential host-interaction functions. As the experimental verification of the effectors is sometimes only based on translocation, the "effector"-status of the proteins with more common functions could be questioned.

More interesting is the group of 290 effectors missing from the outgroup. As could be expected, these effectors include proteins that play an important role in the specific functioning of Legionellales species, Legionella pneumophila in particular. For example, the effector protein VipA changes the cytoskeleton dynamics of the host cell and VipD interferes with the vesicle trafficking by removing a signal protein from the membrane (Qiu and Luo 2017), and they are not found in the outgroup. Furthermore, the effector protein SidE is affecting a multitude of essential functions (Qiu and Luo 2017), such as inhibiting autophagy and regulating ER dynamics, and was also found missing from the outgroup. In addition, the effector RalF has a known function in recruiting essential kinases to the Legionella-containing vacuole (LCV). Since these effectors are essential in the success of Legionellaceae in invading a multitude of hosts, it is logical that these effectors might not be present in the outgroup.

Among the 290 effectors missing from the outgroup were also many proteins with known eukaryotic motives, such as ankyrin repeats and coiled coils. Thus, these genes did not originate from the species in the outgroup. This rules out at least partially, the hypothesis that the eukaryotic-like effectors would have been gained through evolution.
from the common ancestor of these outgroup species and *Legionellales*. They may have, however, originated from other groups outside *Legionellales*. The eukaryotic-like effectors likely play a significant role in adapting to eukaryotic hosts, thus defining the functions in the order *Legionellales*.

Of further interest, are the effectors that can be found both in *Legionellaceae* and the *Coxiellaceae* clades *Aquicella*, *Berkiella* and *Rickettsiella*. There were 38 effector protein clusters present in *Legionellaceae* that were also present in the group *Aquicella*, seven in the group *Rickettsiella* and 22 in the *Berkiella*-group, although they were not present in the outgroup. Not much is known about the effectors present in *Rickettsiella*, except for two of them: one serine/threonine-protein kinase and an ATPase. Among the effector protein clusters in *Aquicella* we could find, for example an ankyrin repeat, a UVB resistance protein and a U-box containing protein. Of these, the ankyrin repeat and the U-box containing protein refer to a link with eukaryotes (de Felipe *et al.* 2005). Similarly to *Legionella*-species, *Aquicella* infect amoeba. This probably explains the acquired or maintained eukaryotic-like effectors, since they are likely essential in the interactions with eukaryotic hosts.

More effector proteins could be identified in the *Berkiella* group. Many of the proteins were shared between all the three species included in the group, likely because two of the species have the same origin. These effectors include an IcmL-like protein, histone methylation motive containing protein, an ankyrin repeat, LigA interaptin, SidB, GTPase activator, RalF, SdeD and a serine/threonine-protein kinase. Of these LigA, SidB, RalF and SdeD are known *L. pneumophila* effectors. LigA has been described as essential for *L. pneumophila*, when infecting its main host, *Acanthamoeba castellanii* (Fettes *et al.* 2000). Further, RalF and SdeD play important roles in high-jacking vesicle trafficking (Qiu and Luo 2017) and ubiquitylation (Luo and Isberg 2004), respectively. Other functions of the effectors found in the *Berkiella* group include interacting with the secretion system (IcmL-like protein) and inflicting potential epigenetic changes in the host cell. In addition, effectors with eukaryotic like domains were found.

The two known *Berkiella* species have been described as invading amoebal nucleus (Mehari *et al.* 2016). Thus, inflicting epigenetic changes in the host may be an essential tool for the *Berkiella* to utilize its host. According to the findings here, they may also hijack other parts of the host cell functioning, such as vesicle trafficking. Further, their secretion may at least partly be similar to that of *L. pneumophila*.

Overall, the distribution of effector proteins in the order reflects previous studies on *Legionella* effector proteins. According to Burstein *et al.* (2016), 38 *Legionella* species share seven core effector genes. The results presented here show a similar conclusion, when six effector protein clusters were found to be common among all *Legionellaceae*, including the newly discovered sequences from TARA marine samples. The distribution of the effector proteins in the other groups further reflects the trend of few shared effectors: in *Aquicella* and *Rickettsiella* the majority of the effector proteins were not present in all.
of the species in the groups. Extremes were seen in the context of Berkiella, where among the species, half of the effectors present were present in all of the species. This may be due to the few number of species in the group so far. On the contrary, the Coxiella-group did not share any of the effector proteins present.

Particularly interesting notation of the effectors present, is the ankyrin repeat protein missing from the outgroup. This protein, with locus tag lpg2300 in \textit{L. pneumophila}, CBU\_1292 in \textit{C. burnetii} and LLO\_0584 in \textit{L. longbeachae} appears even in the group Rickettsiella. It has been previously noted to be conserved among \textit{Legionellaceae} (Burstein \textit{et al.} 2016), and its presence in the other clades has also been noted (Lionel Guy, unpublished).

5.2 Evolution of gene content

A clear concentration of new effector proteins gained in the order could be seen in the earlier branches of the phylogenetic tree (Figures 7, 8). Two events for the most gains could be seen just after the LCA of \textit{Legionellaceae} and in the LCA of the clade containing \textit{L. pneumophila} (Figure 7). As effector proteins are essential in host adaptation, the increase in their number even later in the \textit{Legionellaceae} branches with relatively broad host ranges, is not unexpected. In addition, since most of the effector proteins taken into account in this study originate from \textit{L. pneumophila}, it is expected that many of them have been gained in \textit{Legionellaceae}. However, the effectors that are present, and gained in \textit{Coxiellaceae} show a more regular pattern of a decreasing number of effectors in the later branches of the phylogenetic tree (Figure 8). The patterns of effector gains and losses are thus somewhat distinct between the families. As \textit{Coxiellaceae} tend to be more specialists compared to \textit{Legionellaceae}, the decrease in numbers of new, gained, effectors in \textit{Coxiella} can be expected, and more losses due to adaptations, or loss of selection, in the particular hosts can thus be seen.

According to the results, many of the effector proteins, of which function is known in \textit{L. pneumophila}, were gained at the LCA of \textit{Legionellaceae} or one node after (7). Since the source of most of the verified effector proteins in this study is \textit{L. pneumophila}, it is logical that their origins might be concentrated in \textit{Legionellaceae}. These effectors include LegK1, LegAS4/RomA, LepB, MavN, SidP and RavK. Additionally gained in this node were the locus tags lpg0393 of which function is known, and the aforementioned ankyrin repeat lpg2300. Several of the effectors target cell metabolism and dynamics directly. For example, LegK1 is an eukaryotic like serine/threonine protein kinase, that phosphorylates the NF-\(\kappa\)B inhibitor in the host cell (Rolando and Buchrieser 2014). This activation of NF-\(\kappa\)B inhibits apoptosis, thus allowing the bacteria to escape one of the immune response mechanisms of the eukaryotic cell. This protein is also marked as gained for the clade including Aquicella. However, instead of gaining the protein twice, the more parsimonious hypothesis would be, that it has been lost once in the Coxiella group. In addition, it seems to have been lost from a few individual \textit{Legionella}-species.
Also in the main gained effectors for *Legionellaceae* are LegAS4/RomA. The protein has a methyltransferase activity, methylating histones H3K9/H3K14 (Qiu and Luo 2017, Rolando and Buchrieser 2014). Thus, the effector is capable of inducing epigenetic changes in the host cell. In particular, the effector seems to suppress the host immune system with the methylation (Qiu and Luo 2017).

Further two of the aforementioned group of effectors target the same system regulating vesicle trafficking as the SidE-group of effectors, namely RAB1. These effectors are LepB and the yet unnamed locus tag lpg0393. Thus, hijacking host vesicle trafficking has also evolved fairly early in the evolution of *Legionellaceae*. According to the results from Count, however, these effectors would also have been lost from about 17 *Legionella*-species, including *L. massiliensis*, *L. tunisiensis* and *L. oakridgensis*. All of them seem to be capable of infecting amoeba (Campocasso et al. 2012, Tang et al. 1985), and *L. oakridgensis* is also capable of causing Legionnaire’s disease, although does so rarely. As the pathogen of an eukaryote, *L. oakridgensis*, at least, would benefit in maintaining genes affecting host vesicle trafficking. Thus, it remains unclear, why this loss would have occurred.

The results from Count are further cast in the light of uncertainty due to the predicted late appearance of some of the important effector proteins. Among these are the effectors VipA and SidE. According to the phylogenetic birth-and-death-rates model, both of them would have been gained at the emergence of the clade consisting of *L. longbeachae* and four other *Legionella* species. However, these effectors have been originally annotated and investigated from *L. pneumophila*, which clade separates from the bigger *Legionellaceae*-group before *L. longbeachae*. In addition, according to the gains and losses results, the MavN effector would have been gained also in the *Rickettsiella*-group. This is contrary to our previous results, when looking into the presence of effector proteins in the order, where the MavN effector protein could not be detected in this group (see section 5.1).

Count relies on a probabilistic model for phylogenetic profiles, which is then used in the phylogenetic birth-and-death-rates model (Csur*ó*š 2010). The nature of the probabilities themselves creates a degree of uncertainty to the results. Further, a cutoff of 50% probability was used, which may have been too generous in some cases. Moreover, although the rates themselves are estimated from the data, both individual gene loss rates and gene duplication rates are assumed to be uniform across the members (homologs) of the gene family (Csur*ó*š 2010, Csur*ó*š and Mikl*ó*š 2006). Gene gain by other means, such as horizontal gene transfer (HGT), is treated as a constant (Csur*ó*š 2010, Csur*ó*š and Mikl*ó*š 2006). However, *Legionella*-species seem to readily take up new genes from both closely related species (Gomez-Valero et al. 2011b) as well as from other domains (de Felipe et al. 2005) with HGT. These events would both induce an ”unexpected” increase in the rates for gained genes in the phylogenetic tree. Even though the eukaryotic-like genes have been shown to be fairly conserved among *Legionella*, (Gomez-Valero et al. 2011b), the predictions of gains and losses for the order as a whole may suffer from other parts of the
tree lacking the myriad eukaryotic-like effectors or the prevalence for inter-domain HGT. A hint of this may be seen comparing Figures 7 and 8, where the different dynamics of the two parts of the tree (Legionellaceae and Coxiellaceae) can be seen even based on the predictive models.

Further evidence for HGT may be found from Figures 9 and 10, where many bacteria, unrelated to Legionellales, appear. Such bacteria include Chlamydiales, many enterobacteria and other intracellular bacteria. The trees were based on a pair of core gained Legionellales effector proteins, and could thus indicate, that early in the evolution of Legionellales, a horizontal gene transfer event might have taken place between the ancestor of Legionellales and the other bacteria.

5.3 Early evolution of effector proteins

The core gained effectors were gained at around the last common ancestor of Legionellales, highlighting their importance of the diversification of Legionellales from other bacteria. At around this time, approximately 1.6 billion years ago (Hugoson et al. unpublished), also the distinct type IV secretion system was gained. Thus, the secretion system and the gained effectors at this node are likely to have played an important role in the development of Legionellales ecology as exploiting eukaryotic organisms.

Further looking at the early evolution of the eight gained effector proteins in the LCA of Legionellales, it seems that the origin varied from protein to protein. Two of them had a suggested possible connection to an early ancestor with the eukaryotes (Figures 9 and 10). The other six have an origin in earlier bacterial ancestors, or have been obtained through horizontal gene transfer from other bacteria, as discussed above.

The two clusters with similarities to eukaryotes, an ankyrin repeat (lpg2300) and a Sel1 protein are well known from eukaryotes. The ankyrin repeat areas often interact with cytoskeleton (de Felipe et al. 2005, Gomez-Valero et al. 2011a) and Sel1 is associated with the ER for protein degradation (Mittl and Schneider-Brachert 2007). Both domains are found regularly from bacteria that use amoeba as hosts, as well as from other host associated bacteria, but it has been suggested that these proteins have been gained independently during early evolution (Gomez-Valero et al. 2011a).

Ankyrin repeats, in particular, appear in many organisms, and may be present in many copies. The blast alignment of the ankyrin repeat lpg2300 showed only partial matches with many proteins. Except for some Gammaproteobacteria and Chlamydiales-bacteria, where the alignment matched in full, the proteins partly matched with lpg2300, may consist of non-homologous sections. This casts a veil of uncertainty on the phylogenetic tree constructed from the aligned parts. The mixed grouping of bacteria, fungi and other eukaryotes in the upper branch of the tree (Figure 9) may be a symptom of non-homology or it may reflect the homology of the partial match only. Thus the common ancestry with eukaryotes for this protein is possible, but uncertain based on our results.

Molecular mimicry is in general an important part of host adaptation (Cazalet et al. 24
2004, Gomez-Valero et al. 2011a). According to our results, at least two of the eight gained effectors for the LCA of Legionellales may be linked to adaptation with their current hosts, amoeba, or even earlier protozoa. It may be, that it was these two, the ankyrin repeat lpg2300 (with reservations) and the Sel1 protein lpg0896, that enabled the takeover of the early eukaryotic host by the ancestor of Legionellales. Both of them are still present also in Aquicella. There are, however, many more eukaryotic-like effector proteins present in the order, particularly in Legionellaceae. The choice to concentrate on the particular early nodes of the tree, may thus have limited the extend to which effectors similar to eukaryotic proteins were detected. However, due to the timing of acquirement of these effectors later in evolution or around the LCA of Legionellaceae, their role may have more to do with host adaptation than defining the characteristics of the order.

In the parts where the other six phylogenetic trees can be relied on, it seems that the species in the order Legionellales have been frequently interacting with other bacteria. As previously noted, Legionella are at least prone to transfer genes horizontally among each other (Gomez-Valero et al. 2011b), thus it is plausible that they have been and are readily interacting with other bacteria as well.

6 Conclusion

This thesis project has shown that effector proteins from L. pneumophila, L. longbeacheae and C. burnetii are widely spread among the species of the order Legionellales. However, distinction in the distribution of the effectors could be made between the smaller clades. The investigation into the effectors found in Aquicella did not reveal much more information about the species. This may, however, mean that they have a set of unique effector proteins, that do not have homologs in other Legionellales. Further investigation to these potential genes is thus needed.

Much more was learned about Berkiella-species, by being able to link the function of the effectors to known effectors from L. pneumophila. According to this evidence, many effectors essential for host manipulation are shared between the two clades.

The majority of the effector proteins present in the order seem to have been acquired at fairly early stages of the evolution of Legionellales, confirming what had been proposed before. The core gained effectors at the LCA of the order have thus been significant in the early evolution of the Legionellales. Further huge numbers of genes were gained later in the diversification of Legionellaceae. However, the species in Coxiellaceae seem to have lost more of the studied effectors the closer to present species you go. This may be due to host adaptation in the form of loss of non-essential genes due to loss of selection. As a potential peak of host specificity, the clade Coxiella does not share any of the studied effectors even within the species in the clade, indicating very specialized functions.

Eukaryotic-like effector proteins were found in both Aquicella and Berkiella clades. Possibly some more information about the function of Berkiella-species was uncovered
looking at these effectors, as explained above. Further, some evidence was gained of two core gained effectors having possible links to early eukaryotes, the ankyrin repeat lpg2300 and Sel1 protein family protein. Although a weak evidence, these connections may indicate a horizontal gene transfer event from eukaryotes to Legionellales that has had a particularly important role in distinguishing the order from other gammaproteobacteria, and allowing for the utilization of eukaryotic hosts.

Further evidence of potential HGT from other bacteria was seen from the other core gained effectors. As previously suggested (de Felipe et al. 2005, Gomez-Valero et al. 2014), Legionella-species may be fairly readily taking in genes from other species. This may indicate their flexibility to adapt to a wide host range.

In conclusion, the investigation into effector proteins in the order Legionellales has shown some shared aspects between the species in Legionellales, as well as confirmed some differences between, for example L. pneumophila and C. burnetii. Most prominently, the evolutionary modes seem to differ between the two species. The evolutionary results indicate differences in both evolution regarding host adaptation, as well as in the potential links to eukaryotic ancestors. This study also revealed a set of effector proteins that have likely affected the diversification of the whole order, and were gained in the LCA of Legionellales. With the subsequent gains and losses, the diversification of the clades then took place. Even though some of the effectors are still shared between the clades and give insight into the ecology of the unknown species, more detailed work is needed to study function of the newly discovered species.

7 Acknowledgement

I would like to thank Lionel Guy for offering the interesting topic for the thesis, and for guiding me through the process. My work relied also much on the previous work done by Eric Hugoson. My gratitude goes also to Lisa Klasson for her advice, and to Dennis Leenheer, for sharing the office and indulging my random questions and comments about coding.

8 References


Figure A.1: The workflow of the project methods. Datasets are marked with tilted squares, smaller tasks with hexagons, and other procedures with regular squares. End products are marked with ovals.
Table A.1: Description of the scripts written and used in the project. In addition, the input and output files are described.

<table>
<thead>
<tr>
<th>Script name</th>
<th>Purpose</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>edirect_test.sh</td>
<td>Search all locus tags from NCBI database and fetch protein sequences with annotations</td>
<td>List of effectors</td>
<td>Fasta-files</td>
</tr>
<tr>
<td>combine_info.sh</td>
<td>Adjoin protein accession numbers and annotations to the list of locus tags and their references</td>
<td>Fasta files</td>
<td>csv-file</td>
</tr>
<tr>
<td>protac_to_prodigal.sh</td>
<td>Connect protein accession number to prodigal id and search from OrthoMCL clusters. List then all species and identifiers belonging to the cluster.</td>
<td>Annotations from prodigal to protein accessions. OrthoMCL clustering results.</td>
<td>csv-file</td>
</tr>
<tr>
<td>align_homologs.sh</td>
<td>Combine all sequences from non-unique clusters, align with MAFFT</td>
<td>List of non-unique clusters, Fasta-files</td>
<td>Aligned fasta-files.</td>
</tr>
<tr>
<td>trimm_n_tree.sh</td>
<td>Trim aligned fasta-files with TrimAl and build a phylogenetic tree with IQ-tree</td>
<td>Aligned fasta-files</td>
<td>Trimmed fasta files and .tree-files</td>
</tr>
<tr>
<td>effector_occurrence_forAlleffs.sh</td>
<td>Go through all protein clusters, list all species from them and count how many copies of each cluster are present in each species.</td>
<td>Protein clusters</td>
<td>csv-file</td>
</tr>
<tr>
<td>effector_table.ipynb</td>
<td>Manipulate effector presence data, calculate informative numbers and draw figures</td>
<td>csv-file of effector presence data</td>
<td>various figures</td>
</tr>
<tr>
<td>allModels_GainLoss.sh</td>
<td>List lost and gained protein clusters and corresponding locus tags for each node and count them.</td>
<td>Parsed gain and loss results from Count</td>
<td>csv-file</td>
</tr>
<tr>
<td>effectorSearch_fromGainLoss.sh</td>
<td>List lost and gained effector protein clusters and corresponding locus tags for each node and count them.</td>
<td>Parsed gain and loss results from Count</td>
<td>csv-file</td>
</tr>
<tr>
<td>tree_mods.R</td>
<td>Transfer node annotations (names and numbers) to a list.</td>
<td>an annotated Nexus-tree</td>
<td>txt-file</td>
</tr>
<tr>
<td>tree_visualization_GL.ipynb</td>
<td>Add known annotations to ETE-tree. Import numbers of gained and lost effector protein clusters and visualize on tree.</td>
<td>txt-files with node names and numbers.</td>
<td>Phylogenetic tree with visualisation</td>
</tr>
<tr>
<td>mergeEffectorModels.sh</td>
<td>Combine and align (MAFFT) all sequences belonging to the core gained effector protein clusters</td>
<td>Fasta-files</td>
<td>Aligned fasta-files</td>
</tr>
<tr>
<td>gains_trimm_n_tree.sh</td>
<td>Trim aligned sequences with TrimAl and build phylogenetic trees with IQ-tree</td>
<td>Aligned fasta files</td>
<td>Trimmed fasta-files and .tree-files</td>
</tr>
<tr>
<td>eukaryote_blast.sh</td>
<td>Construct blast-database of each core gained effector protein cluster and run psiblast of each core gained cluster against a database built of itself.</td>
<td>Trimmed fasta-files</td>
<td>pssm-file</td>
</tr>
<tr>
<td>seq_idfix.sh</td>
<td>Modify IDs in fasta-files.</td>
<td>Fasta-files</td>
<td></td>
</tr>
<tr>
<td>blast_hit_trees.sh</td>
<td>Combine protein cluster fasta files with blast hit results, align with MAFFT, trim with TrimAl and build a phylogenetic tree with IQ-tree</td>
<td>Two fasta-files</td>
<td>Aligned and trimmed fasta files and .tree-files</td>
</tr>
</tbody>
</table>
Figure A.2: Modified phylogenetic tree from Figure 1 of the order *Legionellales*, showing the gained (red) and lost (blue) effector proteins predicted by Count for each node. If no gains and losses could be predicted, no graph was added.